

DOCUMENT RESUME

ED 157 905

TH 007 202

AUTHOR Willens, Howard P.; And Others
TITLE Brief Amicus Curiae for Educational Testing Service. United States of America, Plaintiff, and South Carolina Education Association, et al., Plaintiff-Intervenors, v. State of South Carolina, et al., Defendants. In the United States District Court for the District of South Carolina, Columbia Division.
PUB DATE Jul 76
NOTE 185p.; For related document, see TH 007 203
EDRS PRICE MF-\$0.83 HC-\$10.03 Plus Postage.
DESCRIPTORS Constitutional Law; *Court Litigation; Cutting Scores; Effective Teaching; Higher Education; Item Analysis; *Occupational Tests; State Standards; *Teacher Certification; *Teacher Evaluation; Teacher Qualifications; Test Bias; Test Interpretation; Test Reviews; *Test Validity
IDENTIFIERS Educational Testing Service; *National Teacher Examinations; South Carolina; South Carolina Education Association; *USA So Carolina Education Assoc v So Carolina

ABSTRACT

This amicus curiae was filed by the Educational Testing Service (ETS) in the case of the United States of America and South Carolina Education Association v. South Carolina, which involved the validity of South Carolina's use of the National Teacher Examinations (NTE). The NTE, a battery of achievement tests designed to measure the academic preparation of students completing a teacher education program, is produced and administered by ETS. The issues before the court involved the current use of the NTE in South Carolina based upon ETS's study of valid uses of the NTE; the cutting score previously used for certification of prospective teachers in South Carolina; and the use of the NTE for classifying and compensating experienced teachers. Section I of the brief describes the ETS validity study and the action taken by the State Board of Education in reliance upon the study. Section II proposes that the study provides a rational basis for certification decisions and that the State's action in reliance on the study meets the applicable constitutional and statutory standards. Section III evaluates technical criticisms of the study, and concludes that its design and execution both represent informed professional judgment by ETS personnel. The brief, in conclusion, recommends that the court decline to grant injunctive relief; or, if granted, to permit implementation of the certification system adopted by the state in June 1976. (Author/JAC)

 * Reproductions supplied by EDRS are the best that can be made *
 * from the original document. *

PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

E. J. Mason's
Educational Testing
Service

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC) AND
USERS OF THE ERIC SYSTEM "

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRO-
DUCED EXACTLY AS RECEIVED FROM
THE PERSON OR ORGANIZATION ORIGIN-
ATING IT. POINTS OF VIEW OR OPINIONS
STATED DO NOT NECESSARILY REPRESENT
OFFICIAL NATIONAL INSTITUTE OF
EDUCATION POSITION OR POLICY

ED157905

TABLE OF CONTENTS

	Page
Introduction	1
Statement of the Case	2
Interest of Amicus Curiae	3
Argument	6
I. THE NEW USE OF THE NTE ADOPTED BY SOUTH CAROLINA CERTIFICATION AUTHORITIES IS BASED ON A COMPREHENSIVE VALIDITY STUDY THAT WAS CAREFULLY AND PROFESSIONALLY CON- DUCTED	6
A. Background and Purpose	7
B. General Design	9
C. Formation of Panels	14
D. Collection of Data	16
E. Findings	20
F. Action by South Carolina State Board of Education	23
II. THE VALIDITY STUDY PROVIDES A RATIONAL BASIS FOR SOUTH CAROLINA'S NEW CERTIFICA- TION REQUIREMENTS THAT COMPLIES FULLY WITH APPLICABLE CONSTITUTIONAL AND STATU- TORY STANDARDS	26
A. South Carolina's New Certification Rules Meet the Re- quirements of the Fourteenth Amendment	26
1. The Constitutional Standard Requires that South Carolina's Certification System Be Rationally Related to a Legitimate Government Objective	27
2. South Carolina's New Certification Requirements Have the Necessary Rational Basis	28
a. The validity study adequately documents the re- lationship between NTE scores and effective teaching under the relevant professional and legal standards	29
(i) professional standards	29
(ii) legal standards	30

II

TABLE OF CONTENTS—Continued

	Page
b. The use of the validity study by the State Board of Education is rationally related to its objective of providing a sufficient supply of adequate teachers	33
B. The Validity Study Provides Adequate Evidence of Non-discrimination Under Title VII	34
1. Title VII Requires Proof by Defendant that Its Certification Decisions Are Made on a Basis Other than Race	34
2. The ETS Validity Study Provides Adequate Evidence, Under the EEOC Guidelines, of a Non-discriminatory Basis for Certification Decisions	37
a. Requirement with respect to criterion-related study	37
b. Requirement with respect to differential validity study	43
c. Requirements with respect to content validity study	44
d. Plaintiffs' validity study	47
3. South Carolina's Use of the Validity Study Was Proper Under the Guidelines	48
III. THE CRITICISMS OFFERED BY PLAINTIFFS' EXPERTS DO NOT PROVIDE ANY GROUND FOR DISQUALIFYING THE VALIDITY STUDY AS PROVIDING A RATIONAL BASIS FOR CERTIFICATION DECISIONS	52
A. The Design of the Validity Study Was Responsive to the Legitimate Objectives of South Carolina and Represented a Reasonable Exercise of Professional Judgment..	53
1. The Validity Study's Use of a State-wide Standard of Minimal Competence Was Both Reasonable and Necessary	53
a. South Carolina authorities have a legitimate interest in evaluating the academic preparation of graduates from different teacher training programs against a single state-wide standard	54

III

TABLE OF CONTENTS—Continued

	Page
b. The study design was not required to include direct black/white comparisons among institutions	56
c. The NTE can be used properly for the purpose set out in the validity study	58
2. There Was No Necessity to Relate the Content of the NTE to the Job of Teaching by Means of a Formal Job Analysis	60
3. The Use of Teacher Educators Rather than Practicing Teachers on the Knowledge Estimation Panels Was Proper	65
4. The Use of a Majority Rule Standard for Determining Content Appropriateness Was Reasonable	67
5. The Lack of a Pilot Study Provides No Basis for Discarding the Results of the Validity Study	73
B. Plaintiffs' Criticisms of the Execution of the Study Design Involve Alternatives as to which Reasonable Choices Were Made	74
1. The General Quality of Plaintiffs' Evidence Is Such that It Is Entitled to Little Weight	74
a. Testimony of experts	74
b. Testimony of judges	75
(i) timing	77
(ii) leading questions	78
(iii) withholding of documents	79
2. The Criticisms Applicable to Both Panels Are Based Only on Speculation	81
a. Teaching experience of judges at college level....	81
b. Advance preparation by judges	83
c. Use of pre-specified response forms	83
d. Provision of correct answers	84
3. The Criticisms Applicable Only to the Content Review Panels Are Very Limited in Scope and Effect....	85
a. Quality of test content descriptions	85

IV

TABLE OF CONTENTS—Continued

	Page
b. Ability of judges to follow instructions	86
4. The Criticisms Applicable Only to the Knowledge Estimation Panels Have No Foundation in the Record	89
a. Teaching experience of judges at elementary and secondary levels	89
b. Nature of the hypothetical group	90
(i) the dual standard	91
(ii) the different types of hypothetical groups	95
5. The Data from the Judges' Depositions Generally Support the Fairness and Reliability of the ETS Validity Study	100
a. Qualifications of judges	101
b. Willingness of judges to participate	101
c. Absence of any outside influence prior to judges' participation	102
d. Efficient administration of the study	102
e. Clarity of instructions	103
f. Confidence of judges in their ability to achieve fair and reliable judgments	104
C. Plaintiffs' Criticisms of the Data Analysis Are Inconsequential	104
1. Treatment of More/Less/Same Judgments	104
2. Treatment of "Do Not Know" Responses	105
D. The ETS Study Design Was Rigorous, and, if Anything, Created a More Difficult Standard for the Tests to Meet than Was Required Under the Circumstances	108
1. The Expert Opinion on which Plaintiffs' Conclusion Is Based Should Be Viewed With Skepticism	108
2. The Choices Made by ETS Created a Rigorous Standard	110
Conclusion	115
Table 1 Percentages of Questions Judged Content Appropriate by Trial Panels, by Test Edition	21

TABLE OF CONTENTS—Continued

	Page
Table 2 Percentages of Questions Judged Content Appropriate, Not Counting "Do Not Know" Responses, At Various Levels of Agreement Among Judges	69
Table 3 Criteria for Performing Detailed Review of Correspondence between Test Content and Program Content	72
Appendix A: Analysis of Facts and Contentions With Respect to the Common Examinations.	A-1
Appendix B: Analysis of Plaintiffs' Criterion-Related Validity Study of the Use of NTE Scores by South Carolina.	B-1
Appendix C: Analysis of the Judges' Depositions	C-1

VI

TABLE OF AUTHORITIES

CASES:	Page
<i>Albemarle Paper Co. v. Moody</i> , 422 U.S. 405 (1975)	34
<i>Tyler v. Vickery</i> , 517 F.2d 1089 (5th Cir. 1975), <i>cert. denied</i> , 44 U.S.L.W. 3719 (June 14, 1976)	27, 28, 30, 31, 32
<i>United States v. North Carolina</i> , 400 F. Supp. 343 (E.D.N.C. 1975)	1, 4, 5, 8, 27, 28, 47, A-23
<i>Vulcan Soc'y v. C.S.C.</i> , 490 F.2d 387 (2d Cir. 1973)	32
<i>Washington v. Davis</i> , 44 U.S.L.W. 4789 (June 7, 1976)	26, 27, 28, 31, 32
CONSTITUTIONAL AND STATUTORY PROVISIONS:	
U.S. Constitution, Amendment XIV	26
42 U.S.C., § 2000e <i>et seq.</i>	26, 34, 35, 37, 47, 52
ETS PUBLICATIONS:	
Educational Testing Service, GUIDELINES FOR USING THE NA- TIONAL TEACHER EXAMINATIONS (Ex. S-MAN-13)	6, 8
Educational Testing Service, NATIONAL TEACHER EXAMINA- TIONS, TECHNICAL HANDBOOK (1972) (Ex. S-MAN-9)	A-4, A-11, A-17, A-19
Educational Testing Service, REPORT ON A STUDY OF THE USE OF THE NATIONAL TEACHER EXAMINATIONS BY THE STATE OF SOUTH CAROLINA (Jan. 1976) (Ex. S-MAN-32A, S-MAN- 32B)	8, 10, 11, 14, 15, 16, 17, 18, 19, 22, 23, 49, 52, 67, 68, 71, 72, 77, 79, 83, 84, 85, 86, 90, 102, 105, 106, 108, 111, 112, 113, 114
B. Humphry, <i>et al.</i> , ETS AND AACTE PILOT STUDY: TEACHER RATINGS OF KNOWLEDGE OBJECTIVES OF UNDERGRADUATE TEACHER EDUCATION (1975)	A-36
Humphry, <i>Survey of Professional Offerings in NCATE— Accredited Institutions</i> , JOURNAL OF TEACHER EDUCATION 406 (Dec. 1963) (Ex. S-MAN-39)	A-5
B. Pitcher, Relation of Academic Success in Teacher Prepara- tory Curricula with Scores on the National Teachers Com- mon Examination (1962) (Ex. S-MAN-21)	A-21
T. Quirk <i>et al.</i> , A CRITICAL REVIEW OF RESEARCH RELATED TO THE NATIONAL TEACHER EXAMINATIONS (1972) (Ex. P- MAN-8)	39, 49
S. Sherwin, TEACHER EDUCATION: A STATUS REPORT (1974) (Ex. P-MAN-12)	A-13, A-14, A-15
Swineford, An Assessment of the Kuder-Richardson Formula (20) Reliability Estimate for Moderately Speeded Tests (1973)	A-18
N. Wexler, Concurrent Validity of the National Teacher Ex- aminations (1975) (Ex. S-MAN-19)	A-21

VII

TABLE OF AUTHORITIES—Continued

	Page
OTHER REFERENCES:	
American Psychological Association, STANDARDS FOR EDUCATIONAL & PSYCHOLOGICAL TESTS (1974)	10, 11, 29, 30, 36, 41, 42, 62, 63, 64, A-20, A-24
American Psychological Association, STANDARDS FOR PSYCHOLOGICAL TESTS AND MANUALS (1966)	11
Angoff, <i>Scales, Norms and Equivalent Scores</i> , EDUCATION MEASUREMENT 562 (R. Thorndike ed. 1971)	A-16
Bain, <i>Self-Governance Must Come First, Then Accountability</i> , Phi Delta Kappan 413 (April, 1970)	110
Conant, J., THE EDUCATION OF AMERICAN TEACHERS (1963)	A-5
Davis, Relationships Between the National Teacher Examinations, Certain Variables, and Secondary Teacher Education Curricula (unpublished doctoral dissertation 1969)	A-22
Division of Industrial-Organizational Psychology (Division 14), American Psychological Association, PRINCIPLES FOR THE VALIDATION AND USE OF PERSONNEL SELECTION PROCEDURES (1975)	12, 29, 30, 36, 37, 42, 63, 64, A-24
Ebel, <i>The Use of Tests in Educational Licensing, Employment, and Promotion</i> , 8 EDUCATION AND URBAN SOCIETY 19 (1975) (Ex. S-MAN-16)	11
Guidelines For Employee Selection Procedures, 29 C.F.R. Part 1607 (1970)	12, 35, 36, 37, 43, 44, 45, 46, 48, 50, 51, 61, 63
Guion, <i>Content Validity, The Source of My Discontent</i> , — APPLIED PSYCHOLOGICAL MEASUREMENT — (to be published, Fall, 1976) (Ex. P-G-1)	45
H. Gullicksen, THEORY OF MENTAL TESTS (1950)	A-18
Hearings before the Senate Subcommittee on Education on S. 3883 and S. 4167, 91st Cong., 2d Sess. (1972)	110
House, <i>The Price of Productivity: Who Pays?</i> , TODAY'S EDUCATION 65-66 (Sept.-Oct. 1973)	109
Hutchins, <i>Book Review: The Politics of Educational Innovation</i> , 46 HARV. ED. REV. 134 (1974)	109
Koerner, J., THE MISEDUCATION OF TEACHERS (1963)	A-6
Mahaffey, <i>Student Teaching Grades: Assessment and Use</i> , SCHOOL AND SOCIETY 5 (October, 1974)	49
McNeil, et al., <i>The Assessment of Teacher Competence</i> , in SECOND HANDBOOK OF RESEARCH ON TEACHING (Travers ed. 1973)	B-3
Medley, et al., <i>The Application of a Factorial Design to the Study of Cultural Bias in General Cultural Items on the National Teacher Examination</i> , JOURNAL OF EDUCATIONAL MEASUREMENT (Dec. 1974) (Ex. P-MED-12)	A-29, A-30, A-31

VIII

TABLE OF AUTHORITIES—Continued

	Page
Note, <i>Employment Testing: The Aftermath of Griggs v. Duke Power Company</i> , 72 COLUM. L. REV. 900 (1972)	32
D. Ryans, CHARACTERISTICS OF TEACHERS (1962)	B-3
T. Spencer, An Investigation of the National Teacher Examinations for Bias with Respect to Black Candidates (unpublished doctoral dissertation, 1972) (Ex. P-MAN-21)....	A-25, A-26, A-27, A-28
Stanley, <i>Reliability</i> , EDUCATIONAL MEASUREMENT (R. Thorndike ed. 1971)	
Taylor, et al., <i>The Relationship of Validity Coefficients to the Practical Effectiveness of Tests in Selection, Discussion and Tables</i> , 23 JOURNAL OF APPLIED PSYCHOLOGY 565-78 (1939) (Ex. P-MED-15)	50
Wilson, <i>A Second Look At Griggs v. Duke Power Company Ruminations on Job Testing, Discrimination, and The Role of the Federal Courts</i> , 58 VA.L.REV. 844 (1972)	32

IN THE
United States District Court
FOR THE DISTRICT OF SOUTH CAROLINA
Columbia Division

UNITED STATES OF AMERICA,

Plaintiff,

and

SOUTH CAROLINA EDUCATION ASSOCIATION, *et al.*,
Plaintiff-Intervenors,

v.

STATE OF SOUTH CAROLINA, *et al.*,

Defendants.

**BRIEF AMICUS CURIAE FOR
EDUCATIONAL TESTING SERVICE**

INTRODUCTION

This case involves the validity of specified uses by South Carolina of the National Teacher Examinations (hereinafter "NTE"), a series of standardized tests that are produced and administered by the Educational Testing Service ("ETS"), a non-profit corporation located in Princeton, New Jersey. In many respects this case is similar to the companion case brought by plaintiff in North Carolina, *United States v. North Carolina*, 400 F. Supp. 343 (E.D.N.C. 1975) (three-judge court), where the court held, *inter alia*, that a state has a legitimate interest in determining whether persons seeking to teach have the minimum amount of knowledge necessary to teach effectively but that cutoff scores on the NTE can be used to make that determination only if a rational basis is established for such scores. Consistent with this ruling, South Carolina seeks to defend its use of the NTE, including the new certification requirements established on June 11, 1976, on the basis of a professional validity study conducted by ETS.

This brief *amicus curiae* is designed to assist the Court in evaluating the ETS validity study and South Carolina's use of the NTE under applicable professional and legal standards.

STATEMENT OF THE CASE

Plaintiff and plaintiff-intervenors¹ challenge the legality of certain South Carolina laws and regulations which (a) require candidates for certification as teachers in South Carolina to attain prescribed minimum scores on the NTE, (b) allocate the amount of State aid payable to local school districts for teacher salaries in part on the basis of classifications determined by NTE scores of individual educators, and (c) utilize minimum NTE scores on certificates granted on the basis thereof as standards for employing, promoting and compensating teachers. These laws and regulations are challenged principally under the Fourteenth Amendment to the United States Constitution and Title VII of the Civil Rights Act of 1964, as amended, 42 U.S.C. § 2000e *et seq.* In addition to requesting that the pertinent laws and regulations be declared invalid, plaintiffs seek injunctive relief which, *inter alia*, would enjoin the defendants from enforcing these laws and regulations insofar as they require attainment of a prescribed NTE score and from making any employment or certification decision on the basis of such an NTE score.

Pursuant to Order No. 1, proceedings in this case have been bifurcated. The only issue now before the court is whether any of South Carolina's uses of the NTE challenged by the complaint are contrary to federal law. Extensive discovery by the parties has been directed at developing the facts relevant to this central issue. Much of the discovery has focused on documents produced by ETS relating to the NTE generally, the particular uses of the NTE in South Carolina, and the validity study conducted by ETS. Lengthy testimony by an ETS officer and numerous expert and other witnesses was directed to these three general subjects.

The record establishes that South Carolina previously used the NTE with a cut-off score of 975 to decide whether to certify prospective teachers with no experience. It is uncontroverted that proportionately more black persons than white persons failed to

¹ These parties will hereinafter be referred to collectively as "plaintiffs".

achieve this minimum score. In addition, NTE scores have been used by South Carolina as a factor in classifying teachers who fail to achieve the required minimum score. On June 11, 1976, South Carolina established new certification requirements based upon the ETS validity study; the new requirements involve different minimum scores in the various areas of specialization and range between 940 and 1178.

There are, accordingly, three separate issues before this Court relating to South Carolina's uses of the NTE. The first relates to the current certification requirements in effect in South Carolina that were based upon the ETS validity study.² The second involves the minimum score of 975 previously used by South Carolina for certification of prospective teachers; and the third involves the use of the NTE for classifying and compensating some experienced teachers who have not been certified.

INTEREST OF AMICUS CURIAE

In the accompanying Motion for Leave to File Brief as *Amicus Curiae*, ETS describes its interest in this litigation. As the producer of the NTE, ETS has an interest in defending its testing program against a broad scale challenge which appears to question both the integrity and utility of the NTE.³ In addition, ETS has conducted a validity study for South Carolina designed to investigate

² Evaluation by the court of this issue seems both appropriate and desirable. The relevant documents reflecting the action by the State Board of Education are contained in the record as Exhibits S-28 and S-32. Plaintiffs were, of course, fully aware that the ETS study was under consideration by the Board after its delivery in January 1976 and that new certification requirements were likely to be established. In fact, testimony and documents were elicited by the plaintiffs in April 1976 regarding the review of the study by a Board committee and the development of relevant data by the State Department of Education. Maynard deposition, Vol. V, pp. 25-90 (Ex. P-91). In this connection, ETS concurs with the judgment of one of plaintiffs' counsel that the relevance to this lawsuit of the committee's recommendation to the Board regarding new minimum scores "is patently clear." *Id.* at 90.

³ Plaintiffs' suggestion that ETS has a substantial financial interest in selling the NTE (Plaintiffs' Proposed Findings of Fact, ¶ 208(c)) is overblown; income from the NTE program nationwide represented only 2.9% of ETS' total gross income in 1975. Manning deposition, written questions on re-direct examination, Q. 52 (Ex. S-8). In any case, ETS voluntarily determined to sacrifice income from the NTE in South Carolina in 1974 when it decided to refuse to report scores to South Carolina unless changes were made in the use of the NTE.

whether there exists a rational basis for the use of the NTE for certification purposes in South Carolina consistent with applicable legal standards and ETS guidelines. With respect to both of these subjects, ETS believes that this brief *amicus curiae* may assist the court in evaluating the relevant technical issues and professional standards in the measurement field.

As a responsible test producer, ETS seeks to prevent "misuses" of the NTE. South Carolina's use of the NTE for classifying and compensating experienced teachers was considered by ETS to be such a misuse; over a period of several years ETS and South Carolina authorities attempted to resolve their differences amicably so as to permit continued use of the NTE in the State.⁴ When these cooperative efforts failed (largely because of the lack of necessary funding by the Legislature), ETS advised South Carolina that it would no longer report NTE scores to State authorities as of August 31, 1975. As a result of this action and the decision in *United States v. North Carolina, supra*, South Carolina and ETS agreed to undertake a validity study regarding the use of the NTE in the State.⁵

Any general challenge by the plaintiffs to the use of the NTE by South Carolina for measuring the academic preparation of prospective teachers would be foreclosed by the *North Carolina* decision. There the three-judge court held that:

beyond argument . . . the State of North Carolina has the right to adopt academic requirements and written achievement tests designed and validated to disclose the minimum amount of knowledge necessary to effective teaching. 400 F. Supp. at 348.

The court also concluded that the NTE:

do not measure teaching skills but do measure the content of the academic preparation of prospective teachers. We are inclined to think that the NTE tests measure the critical mass of knowledge in academic subject matter and that a score somewhere on the scale would disclose the knowledge necessary as a prerequisite to effective teaching. *Id.* at 350.

⁴ See Stipulation No. 6, Attachments SM-36 through SM-90 (Ex. P-117).

⁵ Further action by South Carolina to bring its practices into full compliance with ETS guidelines is awaiting this court's determination of the issues raised by the complaint.

Notwithstanding this judicial assessment of the NTE, plaintiffs renewed their general attack on the NTE during the discovery in this proceeding with particular reference to the Common Examinations. Although these familiar contentions regarding the NTE have apparently been abandoned by plaintiffs, ETS believes that the court should be fully satisfied regarding the professionalism with which the NTE are developed and administered. Accordingly, Appendix A to this brief contains a full description of the facts with respect to the purpose, construction, scoring, and statistical properties of the Common Examinations and deals with the various criticisms of the tests offered by plaintiffs' witnesses.

The new issue in this litigation which distinguishes it from the *North Carolina* case arises from the fact that South Carolina seeks in part to defend both its past and present uses of the NTE by a professionally conducted validity study. In *United States v. North Carolina*, the court held that the minimum score on the NTE required by North Carolina violated the Fourteenth Amendment because it lacked any rational basis. The court stated there that, although a cut-off "score somewhere on the scale would disclose the knowledge necessary as a prerequisite to effective teaching," where "that point is—whether at 950 or some other score—is not established by the record." 400 F. Supp. at 350. In this case, on the other hand, a full record does exist with respect to the basis upon which South Carolina recently established minimum NTE scores for certification, most particularly the ETS validity study and its consideration by South Carolina authorities in light of their educational objectives. The adequacy of this record under the "rational basis" Constitutional test of the *North Carolina* case and under Title VII of the Civil Rights Act of 1964 (assuming the Constitutional test is met) is thus squarely before this court. It is this issue which ETS will address in this brief *amicus curiae*.

ETS does not address in this brief plaintiffs' request for declaratory or injunctive relief with respect to the State's past use of the NTE for certification purposes.⁶ ETS believes that the aspects of this issue that are relevant to the validity study would be decided on the basis of the same facts as are discussed in this brief, and that other questions with respect to the State's past use,

⁶ ETS does not separately address the use of the NTE by school districts for selection purposes since those uses are largely derivative of the certification use.

such as those relating to alleged discriminatory intent or the existence of a rational basis for the State's actions apart from the ETS validity study, are beyond ETS' role as an *amicus*. With respect to injunctive relief, all but one of the score levels recently adopted by the State are substantially above the 975 previously required by the State for certification. If, as ETS believes, the validity study provides a rational basis for those higher cut-off scores, then the issues with respect to injunctive relief directed at the lower score are largely academic. Injunctive relief, if appropriate, would be directed at the certification system currently in effect that uses new score levels based on the ETS study.

ETS does not address the issues with respect to monetary relief arising out of the use of the NTE to determine salary levels. ETS' position on this use has been published in its GUIDELINES FOR USING THE NATIONAL TEACHERS EXAMINATIONS (1974) (Ex. S-MAN-13).

ETS has organized this brief *amicus curiae* as follows: Section I describes the validity study and the action taken by the South Carolina State Board of Education in reliance upon the study. Section II demonstrates that the validity study provides a rational basis for certification decisions and that the State's actions in reliance on the study meet the relevant Constitutional and statutory standards. Section III rebuts the various criticisms offered by plaintiffs in their attempt to disqualify the validity study as a basis for the new certification system. Most of these criticisms had been anticipated and evaluated in advance by ETS, and decisions made by ETS with respect to matters of design, execution and data analysis represent an informed exercise of professional judgment fully consistent with applicable professional and legal standards.

I. THE NEW USE OF THE NTE ADOPTED BY SOUTH CAROLINA CERTIFICATION AUTHORITIES IS BASED ON A COMPREHENSIVE VALIDITY STUDY THAT WAS CAREFULLY AND PROFESSIONALLY CONDUCTED.

Central to the issues before this court is the validity study conducted by ETS for the South Carolina Department of Education (Department). Dated January 1, 1976, this unprecedented validity study was designed to explore whether the use of the NTE in South Carolina could be validated. Drawing upon the professional resources of ETS and the expertise of four independent reviewers, the study

collected and analyzed the individual judgments of 456 experienced educators representing 25 teacher training institutions in South Carolina. The study concluded that the overwhelming majority of questions on the NTE (86% on the Common Examinations and more than 90% on 13 of the 20 Area Examinations studied) are appropriate for use in South Carolina. It proposed minimum NTE scores for use in evaluating whether prospective teachers in South Carolina possess the desired level of academic knowledge. In this section of the brief, ETS will summarize the background and purpose of the study, explain its design and operation, summarize its findings, and describe the new certification requirements established by the State Board of Education in reliance upon the study.

A. Background and Purpose

The NTE are a battery of achievement tests produced and administered by ETS that are designed to measure the academic preparation of students completing a teacher preparation program. The NTE have two principal components, the Common Examinations, which test elements of teacher training common to all fields, and the 28 Area Examinations, each of which tests knowledge specifically relevant to a specialized area in teaching or school administration. The Department has used the NTE for many years to measure the academic preparation of candidates seeking to teach in the State. In 1974, ETS notified the Department that scores on the NTE would no longer be reported to it as of August 31, 1975. ETS took this action because NTE scores were being used in South Carolina as a basis for classifying and compensating teachers who often held extensive records of job performance within the State.⁷ Such use of NTE scores violated the publicly-stated position of ETS

⁷ The correspondence and other documents describing this difference of views and the efforts of both parties to reach an acceptable resolution is set forth as part of Stipulation No. 6, Attachments SM-36 through SM-90 (Ex. P-117).

The parties have devised a citation system for exhibits that are part of the record. Those citations include letter and numerical designations that are lengthy, confusing and difficult to use because most of the designations used during the depositions, in which most of these exhibits were identified, have been changed. For this reason, ETS has used a citation form that describes each document or gives its title and includes thereafter the exhibit number apparently designated by the clerk. ETS has used a list of exhibit numbers provided by plaintiffs.

that the scores should not be used to determine teacher salaries or to classify experienced classroom teachers."

Shortly before August 31, 1975, ETS and Department representatives agreed on a resolution of their differences that would permit continued use of the NTE by South Carolina consistent with ETS guidelines.⁹ As part of this accommodation, the Department and ETS agreed that a study would be conducted regarding the relationship of the NTE to teacher training programs in South Carolina and the use of NTE minimum scores for purposes of initial certification.¹⁰ On October 7, 1975, ETS submitted a proposal to the Department, including a detailed Statement of Work.¹¹ The proposal was accepted formally by the Superintendent of Education on October 29, 1975. The study was conducted by ETS during October, November and December and the final report was delivered to South Carolina on January 7, 1976.

The study had two specific purposes: (1) to assess the appropriateness of the NTE as a measure of knowledge acquired through teacher preparation programs and required as a qualification to teach in South Carolina; and (2) to develop information that State officials could consider in establishing rational rules for evaluating the NTE scores of candidates for initial certification. To achieve these purposes, the study was designed to address both (1) the relationship between the content of the NTE and the content of teacher training programs within the State; and (2) the level of performance on the NTE that might be expected of a candidate

⁹ The NTE are prepared for a limited well-defined purpose—to measure the preparation of prospective teachers against a national norm. Accordingly, ETS has defined certain uses of the NTE that are consistent with their purpose and other uses that are not. GUIDELINES FOR USING THE NATIONAL TEACHER EXAMINATIONS (1974) (Ex. S-MAN-13).

⁹ These discussions are summarized in the letter dated August 19, 1975, from Cyril B. Busbee, State Superintendent of Education, to Dr. Winton H. Manning, Vice President of ETS (Ex. P-M-10) and Dr. Manning's letter to Dr. Busbee dated August 25, 1975 (Ex. P-M-11).

¹⁰ By this date, of course, the South Carolina authorities were aware of the decision in *United States v. North Carolina*, 400 F.Supp. 343 (E.D.N.C. 1975), that the use of minimum scores on the NTE without any rational basis for such scores was unconstitutional.

¹¹ The Statement of Work was published as part of the ETS report of the study, entitled REPORT ON A STUDY OF THE USE OF THE NATIONAL TEACHER EXAMINATIONS BY THE STATE OF SOUTH CAROLINA (hereinafter the "REPORT"), at pp. B-1 through B-11 (Ex. S-MAN-32A).

who is judged to have the minimum amount of knowledge necessary to complete an academic program in teacher education in South Carolina and to teach effectively in South Carolina schools.

B. General Design

Although ETS carries out hundreds of validity studies each year, it recognized that the proposed South Carolina study required a specially designed approach that would satisfy all pertinent professional standards, meet the particular needs of the Department, and withstand critical scrutiny by the parties to this litigation and the court. The project was accordingly assigned the highest priority by ETS management and an outstanding team of ETS specialists was assembled to design and conduct the study. The team of 18 ETS professionals was supervised by two ETS Vice Presidents—Dr. Winton H. Manning and Dr. Richard L. Burns.¹² The ETS project staff was co-directed by Dr. E. Elizabeth Stewart and Dr. Edward Masonis, with Dr. Norman Wexler serving as project statistical coordinator. An internal advisory committee was formed which included technical experts with specialized skills in the fields of test development, statistics, measurement, data interpretation, and test validation.¹³ Plaintiffs have not questioned the professional qualifications of the ETS team responsible for the validity study.

After careful consideration of the dimensions of the task, the ETS team decided that the "content validity" method should be used in the study. This selection of methodology seemed to comport best with the relevant professional standards and the particular goals of this validity study. Evidence of content validity consists of showing that what a test measures constitutes "a representative sample of behaviors to be exhibited in a desired performance domain"; "[c]ontent validity is most commonly evaluated for tests of skill or

¹² The qualifications of Dr. Manning are set forth in detail in his deposition at pp. 16-23 (Ex. S-8) and are supplemented by a list of his publications that is Exhibit S-MAN-1. The qualifications of Dr. Burns are set out in Exhibit S-MAN-33.

¹³ The curriculum vitae of the ETS staff members who contributed to the South Carolina validity study are contained in Exhibit S-MAN-33. These documents set forth the educational credentials of the ETS personnel, their employment histories, publications, professional affiliations, and professional honors.

knowledge," i.e., achievement tests.¹⁴ Since the NTE are achievement tests designed to measure academic preparation and do not purport to predict effectiveness in the job of teaching, the content validation approach was selected. The principal task for the ETS research team was the development of a specific study design that applied the general methodology to the special challenges of the South Carolina assignment.

The comparison of the content of the NTE and the teacher education curriculum throughout South Carolina required a study design that permitted both very specific comparisons and a relatively broad overview. In addition, the task of assessing the level of performance to be expected of a minimally qualified candidate for certification required access to a broad range of experience with teacher preparation and teaching in South Carolina. In order to meet these requirements, ETS concluded that the most appropriate design for the study would be one which systematically obtained, recorded and analyzed the individual judgments of a large number of educators with extensive experience in the preparation of teachers in South Carolina. The educators who participated in the study were grouped into panels or juries to consider specific parts of the overall comparisons and assessments required by the design.

Before deciding finally upon this particular study design, ETS explored possible alternatives to the jury judgment approach. For example, ETS considered the feasibility of a systematic content analysis of written documents at each of the 25 teacher training institutions in South Carolina. Theoretically, at least, such an examination would permit comparison of the material taught in the teacher training programs with the knowledge tested by the NTE. Upon further analysis, however, it became obvious that this alternative would be excessively time-consuming, would require extensive manpower, and would still provide no meaningful guarantees that the actual content of the curriculum had been accurately and completely surveyed.¹⁵

The study design developed by ETS draws upon a body of professional literature suggesting the appropriate uses and possible

¹⁴ American Psychological Association, STANDARDS FOR EDUCATIONAL & PSYCHOLOGICAL TESTS at 28 (1974) (hereinafter "APA STANDARDS") (Ex. P-MAN-3).

¹⁵ The reasons why this and another alternative to the jury judgment approach were rejected are summarized in the REPORT at pp. 9-11 (Ex. S-MAN-32B). A truncated analysis of this sort is set out in Appendix E to Plaintiffs' Proposed Findings of Fact without any evidentiary foundation.

limitations of the jury judgment method. Although none of these earlier studies involved the use of jury judgments on such a large scale, the studies revealed some of the problems that might arise. With these precedents in mind, the ETS team modified the study design to provide appropriate procedures and safeguards to maximize the accuracy and usefulness of the proposed study.¹⁶

As part of its refinement of the study design, ETS solicited the views of four highly qualified experts not affiliated with ETS. Each was asked to comment upon the proposed study design, to advance any criticisms (real or fancied) which he thought might be directed at the study, and to make suggestions for improvement. The four experts were:

Robert L. Ebel: Dr. Ebel received his Ph.D. in educational psychology at the University of Iowa in 1947 and is a professor of education and psychology at Michigan State University. He was an ETS vice president from 1957 to 1963. Dr. Ebel is a past president of the APA Division of Evaluation and Measurement (1971-72) and a past president of the American Educational Research Association (1972-73). He has written extensively on the relevance of knowledge to effective teaching and is concerned with problems of teacher certification.¹⁷

Robert M. Guion: Dr. Guion received his Ph.D. in industrial psychology from Purdue University in 1952 and is currently a professor of psychology at Bowling Green State University. He is a former president of Division 14 of the American Psychological Association (1972-73). Dr. Guion is the principal author of the APA STANDARDS¹⁸ and one of the authors of the 1966 predecessor publication, the STANDARDS FOR PSYCHOLOGICAL TESTS AND MANUALS. Because of his "innu-

¹⁶ For example, some of the earlier studies evaluated the ability of the judges to estimate the difficulty of questions for students depending upon whether the judges had information regarding the actual difficulty of similar questions as measured by student performance. The experience reflected in these studies suggested to ETS that the judges in the proposed South Carolina study should have some exposure to actual performance on NTE questions before being asked to estimate the percentage of minimally qualified candidates for initial certification in South Carolina who would be expected to know the answers to particular NTE questions. As a result, ETS devised a training exercise for the judges which is discussed in the REPORT at pp. 53-56, 61.

¹⁷ See, e.g., Ebel, *The Use of Tests in Educational Licensing, Employment, and Promotion*, 8 EDUCATION AND URBAN SOCIETY 19 (1975) (Ex. S-MAN-16). Dr. Ebel's qualifications are further set out in his curriculum vita which is Exhibit S-MAN-17.

¹⁸ See APA STANDARDS, Foreword; Guion deposition, p. 20 (Ex. S-5). Dr. Guion's qualifications are set out in detail in Exhibits S-G-1 and S-MAN-34.

merable" qualifications, Dr. Guion was also designated by Division 14 as chairman of the drafting committee and principal co-author of the PRINCIPLES FOR THE VALIDATION AND USE OF PERSONNEL SELECTION PROCEDURES.¹⁹ In addition, Dr. Guion has served on the advisory committee that drafted the current Guidelines on Employee Selection Procedures of the Equal Employment Opportunity Commission,²⁰ 29 C.F.R. Part 1607, and was a participant in drafting the Uniform Guidelines on Employee Selection Procedures, now being prepared by the Equal Employment Opportunity Coordinating Council.²¹ He was described by plaintiffs' most eminent expert as a "highly regarded personnel psychologist who has written a text in personnel psychology and who has taken fairly major responsibility for the testing guidelines of the American Psychological Association. My impression is that he is very concerned about and sensitive to the problems of equity in testing for minority groups."²²

David R. Krathwohl: Dr. Krathwohl received his Ph.D. in educational and psychological measurement from the University of Chicago in 1953 and is the former dean of and a professor in the School of Education at Syracuse University. He is a past president of the American Educational Research Association (1968-69) and a past president of the APA (1973-74). Dr. Krathwohl is directly concerned with, and is an expert in, the field of the academic preparation of teachers.²³

Robert L. Linn: Dr. Linn received his Ph.D. in measurement from the University of Illinois in 1965 and is currently a professor of educational psychology and psychology at that University. He was formerly a senior research psychologist at ETS. Dr. Linn is a past chairman of the APA Committee on Psychological Tests and Assessment (1975), and has served as an associate editor of the JOURNAL OF EDUCATIONAL STATISTICS as well as an advisory editor of both the JOURNAL OF EDUCATIONAL PSYCHOLOGY and the JOURNAL OF EDUCATIONAL MEASUREMENT. He has published a substantial number of articles dealing with issues of fair testing, and was described by plain-

¹⁹ Division of Industrial-Organizational Psychology, American Psychological Association, PRINCIPLES FOR THE VALIDATION AND USE OF PERSONNEL SELECTION PROCEDURES, Foreword (1975) (hereinafter "DIVISION 14 PRINCIPLES").

²⁰ Guion deposition, pp. 16-17 (Ex. S-5).

²¹ *Id.* at 21.

²² Thorndike deposition, p. 86 (Ex. P-110).

²³ Dr. Krathwohl's qualifications are set out in Exhibit S-MAN-34.

iffs' expert as having "a particular concern for problems of fair use of tests in relation to minority groups."²⁴

Each of the four experts submitted detailed comments to ETS.²⁵

With respect to the overall study design, the independent reviewers were uniformly complimentary. Dr. Linn expressed his view that the study

would represent a real step forward, not only for the NTE but for the process of content validation. Overall the proposal is of excellent technical quality.²⁶

Similarly, Dr. Guion described the study as being "a pioneering kind of design" which, using "a very creative method," did a "thorough job in laying out the issues that have to be considered and establishing an operating cutting-off score, which is a policy decision, not a technical one."²⁷ When asked whether the design was "professionally acceptable," Dr. Guion responded:

I thought that it went beyond what would be called 'professionally acceptable'. . . I would characterize as 'professionally acceptable,' in quotes, as being something minimally acceptable or something of that sort. And this would certainly be more than that.²⁸

Dr. Ebel and Dr. Krathwohl also responded favorably to the proposed study design, finding it to be professionally acceptable in all respects.²⁹

As they were requested to do, however, the four experts gave ETS their views regarding how the study might be criticized and how it might be improved. All their comments were carefully reviewed by ETS. Many of the suggestions proposed alternate pro-

²⁴ Thorndike deposition, pp. 87, 92 (Ex. P-110). Dr. Linn's qualifications are set out in Exhibit S-MAN-34, and in his deposition at pp. 5-14, 137-45 (Ex. P-116).

²⁵ Letter from Linn to Stewart, dated Sept. 27, 1975; letter from Linn to Stewart, dated Nov. 24, 1975 (P-MAN-18); letter from Krathwohl to Masonis dated Oct. 5, 1975; letter from Krathwohl to Masonis dated Oct. 27, 1975 (Ex. P-MAN-19); Guion deposition, pp. 23-26 (Ex. S-5); Manning deposition, pp. 260-63, 893-922 (Ex. S-8).

²⁶ Letter from Linn to Stewart, dated Sept. 27, 1975 (Ex. P-MAN-18).

²⁷ Guion deposition, pp. 24, 34 (Ex. S-5).

²⁸ *Id.* at 24.

²⁹ Manning deposition, pp. 260-63 (Ex. S-8).

cedures that would address the questions considered in the study in a different but not necessarily preferable way and others identified additional questions that might be addressed as part of the study. Where the comments were consistent with the overall methodology and could be accommodated within the general constraints set out by South Carolina, ETS did make changes in response to some of the suggestions of the four reviewers. All the suggestions were evaluated by ETS and sound reasons exist for the decisions of the ETS professional team with respect to each of them.³⁰

C. Formation of Panels

The study design called for the creation of two types of panels, each with a separate assignment. The Content Review Panels evaluated the congruence between the content of the NTE and the curriculum prescribed for teacher education programs in the State. The Knowledge Estimation Panels estimated the percentage of minimally knowledgeable candidates for initial certification in South Carolina who would be expected to know the answers to the questions on the NTE. The selection of educators to serve on these panels and the formation of representative panels were central to the success of the validity study.

ETS determined the number and size of the panels after weighing several relevant considerations. Because of the different subject matters covered by the four tests in the Common Examinations and the 21 Area Examinations used by the State, it was generally decided that each test would be considered by separate panels.³¹ Although it would have been economical to ask the same faculty

³⁰ All of the suggestions made by the reviewers and the ETS decisions with respect to these suggestions are set out in Manning deposition, written questions on re-direct examination, Q.11-28 (Ex. S-8).

³¹ Although it was originally planned to have panels for each of the 21 separate Area Examinations required by the State for teacher certification, eventually the test in Teaching Hearing Handicapped was dropped from the study because only one faculty member was available to serve as a judge for it. In addition, exceptions were made for two of the Common Examinations which covered such disparate subjects that it was considered unrealistic to expect to locate educators whose backgrounds would qualify them to make judgments about many of the questions in both parts of the tests. Two panels of each type instead of just one were created for each of those tests. REPORT, pp. 21-22 (Ex. S-MAN-32B).

members to serve on both the Content Review Panel and the Knowledge Estimation Panel in their area of expertise, ETS concluded that such multiple assignments would complicate the judge's task and might lessen the accuracy of their judgments. With respect to the size of the panels, it was necessary to consider the probable availability of faculty members in a given subject area, the need to represent appropriately the different teacher training institutions, the need to match the backgrounds of the panel members to the content coverage of the test, and the need to be able to divide as many panels as possible into comparable halves for analytical purposes. With these considerations in mind, ETS decided to include a minimum of 10 faculty members on each panel whenever this number of qualified judges was available.³²

The composition of the panels also reflected ETS' efforts to obtain qualified educators who could fairly represent each of the 25 teacher training institutions in South Carolina including those with significant numbers of minority group students. Based upon information supplied by the Department, ETS calculated the relative size of the programs offered at the different institutions in fields where a test was required for certification, and the proportions of 1973-74 graduates from these programs who had attended institutions at which the student body is predominantly black or predominantly white. Information regarding the number of institutions and graduates was used to determine the total number of nominations to be requested in a particular field, although it was also necessary in this connection to consider the somewhat complex relations between the subject matter areas covered by the various tests and the organizational structures of the institutions.³³ It was decided that the teacher training institutions themselves should nominate faculty members to participate in the study in order to avoid any predisposition favorable to the NTE.

After the various determinations as to the panel requirements were made by ETS, the Department of Education solicited the 25 institutions to secure the desired number of nominees from each

³² Larger minimum sizes were specified for three parts of the Common Examinations and for five Area Examinations. REPORT, pp. 23-24 (Ex. S-MAN-32B).

³³ As discussed in the study, faculty members in some subject areas provided a pool for selecting judges for various tests. Also, the panel for some tests included faculty members from several different areas. In one instance faculty members in six traditionally distinct fields of science were needed, in different mixes, for the panels for each of three different tests. *Id.* at 24-25.

institution to participate in the study. To ensure that all panel members would have a reasonable amount of relevant teaching experience, the institutions were asked to nominate only persons who had taught in South Carolina colleges for at least two years. The institutions were asked to describe the race, sex and teaching emphasis of each nominee and to state the nominee's number of years of teaching experience in South Carolina. As reflected in the study, the 747 nominees obtained through this process were generally representative of the 25 institutions.³⁴ To some extent, the smaller institutions nominated more faculty members in relation to their size than did the larger institutions. In addition, predominantly black institutions were slightly overrepresented among the nominees, and state supported institutions were somewhat underrepresented in relation to privately supported institutions.

From the original group of 747 nominees, ETS invited 515 to serve as judges on one of the panels for a particular test. In making these selections, several factors were given priority. First, to the extent that the pool of nominations permitted, the prespecified numerical representation of predominantly black and predominantly white institutions was effected.³⁵ Second, an attempt was made to ensure that the panel members would provide appropriate coverage of the subject matters covered by the test. Third, an attempt was made to select panel members so as to have a fair representation of the varying sizes of institutions providing teacher training programs in the State. Of the 515 persons initially invited to serve on panels, 84% actually did so. After appropriate replacements had been obtained for those who were unable to participate, a total of 456 judges participated in the panels and contributed the judgments which were the essential raw data of the validity study.

D. Collection of Data

After the panel members were selected, ETS sent them preliminary materials to assist in their preparation. Members of the Content Review Panels received materials describing their assignment, containing instructions with respect to their review of either

³⁴ REPORT, pp. 29-35 (Ex. S-MAN-32B).

³⁵ Except when an insufficient number of nominees was available, each panel included at least two representatives of schools attended predominantly by black students. *Id.* at 35.

the Common Examinations or the Area Examinations, and describing the content of the various tests. Members of the Knowledge Estimation Panels received comparable materials. Members of all the panels received the various forms prepared by ETS on which their judgments were to be recorded.³⁶

The panels of educators met at three different locations in South Carolina. Each assembly of panel members was staffed by appropriate ETS personnel and a representative of the Department of Education. At the orientation session the procedures for the study were generally reviewed, further oral instructions were given, and the panel members were encouraged to ask questions about their assignments. Subsequently the panels assembled and began the process of making their judgments regarding the NTE. Although the educators were assembled as a group, the instructions specifically provided that their judgments were to be made individually and independently; members of the same panel did not confer as a group and no member of a panel was informed of the judgment made by any other individual member. It was also made clear that the judgments were made in confidence and that no participant would be publicly identified with his responses. These procedural safeguards were designed to ensure the independence of the panel member's judgment.

The Content Review Panels were asked to make four independent types of judgments about the congruence of the NTE tests they were reviewing and the curriculum in teacher training programs in South Carolina. Judges were asked to indicate, on the

³⁶ The materials sent to the panel members are set forth in the REPORT at pp. E-1 through E-44 and G-1 through G-13 (Ex. S-MAN-32A), and are discussed in the REPORT at pp. 56-57 (Ex. S-MAN-32B).

The members of the Knowledge Estimation Panels received additional preparation in order to assist them in estimating the percentages of minimally qualified examinees who would know the correct answer to particular NTE questions. ETS developed a training exercise which contained a set of sample questions accompanied by an estimate of the percentage of examinees at each of several score levels who would know the correct answer. These data were based on a group of NTE examinees very similar in overall score levels to all South Carolina seniors who take the NTE for the first time. Before the judges were told the percentages of examinees actually answering the questions correctly, they were given an opportunity to make their own estimates of the percentages of students in each grouping who would know the answer. The exercise was designed to supply some guidance to the judges regarding actual performance on the NTE before they were asked to estimate the performance of the hypothetical group of minimally qualified candidates. *Id.* at 53-56, 61.

basis of a review of the major content topics of the test, whether the emphasis in the curriculum on each content topic was about the same as, more than, or less than the percentage of the test devoted to the content topic. As an additional means of assessing the overall similarity between tests and curricula, judges were also asked to indicate one of four characterizations of greater or lesser similarity between the test and the curriculum of their institution.³⁷ As a measure of the comprehensiveness of the tests, judges were asked to indicate which, if any, major content topics in the teacher education sequence were not included among the test content topics. Finally, the judges were asked to review each question on the test to determine whether the individual question was appropriate for use in South Carolina. The questions on the NTE were regarded as "content appropriate" if at least 51% of the judges indicated that at least 90% of the students taking the test would have had an opportunity to learn the answer as part of their teacher education program or in courses prerequisite to entering it.³⁸

The task of the Knowledge Estimation Panels was to estimate the percentage of minimally knowledgeable students who would know the answer to each individual test question. Based on their experience, members of the Knowledge Estimation Panels were individually required to create a hypothetical group of persons whom they would judge to have the minimum amount of academic knowl-

³⁷ The judges were instructed:

Based on the information you have regarding the general content topics covered in this test, select the option below that most closely characterizes your judgment regarding the similarity between this NTE test and the teacher education sequence at your institution.

- A The test content topics parallel the teacher education sequence at our institution very closely.
- B There are some differences between the test content topics and the teacher education sequence at our institution, but these differences do not appear to be appreciable.
- C There appear to be some appreciable differences between the test content topics and the teacher education sequence at our institution.
- D There is little similarity between the test content topics and the teacher education sequence at our institution.

REPORT, p. E-3 (Ex. S-MAN-32A).

³⁸ The 90% figure was selected on the basis of professional judgment and was believed by ETS to provide a very conservative standard for classifying the test questions as to appropriateness.

edge needed to complete the teacher education program required for certification in South Carolina and the minimum amount of knowledge to teach effectively.³⁹ Second, these panel members were asked to estimate the percentage of this hypothetical group who would know the correct answer to a particular question (without taking into account the possibility that some additional students might pick the correct answer purely as a lucky guess).

The completed response forms were analyzed by a computer program developed especially for this purpose. As a quality control step, parallel data processing by clerks was performed on a limited basis. Before the data were analyzed, ETS made a few necessary adjustments to deal with unforeseen difficulties which had arisen during the data collection phase. If any doubt developed regarding the reliability of the data with respect to a particular question reviewed by the judges, ETS eliminated the question from the analysis.⁴⁰ The data from the Content Review Panels were analyzed first. Only questions that were determined by a majority of the judges to be content appropriate for use in South Carolina were considered in the next phase of the data analysis, which was to summarize the judgments of the Knowledge Estimation Panels with respect to the percentage of minimally qualified students who would answer the question correctly. The estimates made by the individual judges were combined statistically and these combined estimates were then analyzed to generate scaled scores that reflected for each test the level that would be achieved by a minimally knowledgeable candidate. The consistency of the results was analyzed in two ways: ETS evaluated the comparability of the results for two different test editions and, when the number of judges permitted, the degree of similarity between the judgments of half panels.

³⁹ In theory, the minimum amount of knowledge to graduate from an approved program should be the same as the minimum amount of knowledge to teach effectively. In practice, it seemed likely that the quality of the programs at the 25 teaching training institutions in South Carolina differed sufficiently that those standards might not be the same. The judges were asked to use *both* standards. The practical effect of this instruction was to require each judge to make a determination as to which standard was *higher* (with respect to the curriculum at his or her institution) and to apply that higher standard.

⁴⁰ The processing of the data and the adjustments required are discussed in the REPORT at pp. 64-66. (Ex. S-MAN-32B).

E. Findings

The major findings with respect to the content review of the NTE in relation to the teacher education curriculum in South Carolina were as follows:

1. For the Common Examinations 86% of the questions were classified as content appropriate; for four of the six subparts of the Commons the figure exceeded 90%. For 13 of the 20 Area Examinations studied, at least 90% of the questions were classified as content appropriate. These findings are set out in Table 1.

2. Almost four-fifths of the judges named no major content area in the curriculum that does not appear in the descriptions of test content. Only 11 topics, with respect to eight tests, were mentioned by more than one judge apiece.

3. Judged differences in the relative emphasis placed on specified topics in the tests and the teacher training curriculum were summarized by an index with values ranging from 0 (representing identity of relative emphasis) to 100 (representing maximum possible divergence). The value of the index was .32 for the entire set of Common Examinations and below 40 for all but three of the Area Examinations.

4. Overall, 65% of the judges reported either close parallelism or no appreciable differences between the content of the tests and the teacher education sequence. Such an evaluation was submitted by at least half the judges for 15 of the 20 Area Examinations and by a weighted averaged of 69% of the judges for the Common Examinations.

5. A comparison of three of the indicators of the relation between the tests and the teacher education programs showed that the variations with respect to specific tests were generally consistent in direction. Analysis of the results for agreement between test editions and between half panels demonstrated a relatively high level of consistency.

TABLE 1

Percentages of Questions Judged Content Appropriate by
Total Panels, by Test Edition

Examination	Edition 1		Edition 2		Average ^b %
	n ^a	%	n	%	
Common Examinations		87.4 ^c		84.8 ^c	86.1
Professional Education	109	86.2	110	80.0	83.1
Written Eng. Expression	45	100.0	45	93.3	96.6
Social Studies	29	96.6	30	100.0	98.3
Literature & Fine Arts	33	60.6	35	65.7	63.2
Science	30	93.3	30	90.0	91.6
Mathematics	20	90.0	20	95.0	92.5
Area Examinations					
Agriculture	99	96.0	73	95.9	96.0
Art Education	150	71.3	150	74.7	73.0
Biology & Gen. Sci.	159	96.2	159	93.7	95.0
Business Education	153	93.5	160	98.8	96.2
Chem., Phys., & Gen. Sci.	149	93.3	150	94.7	94.0
Early Childhood Education	125	98.4	122	94.3	96.4
Educ. in the Elem. School	148	90.5	150	92.0	91.2
Educ. of the Ment. Ret.	149	98.7	150	94.7	96.7
Eng. Lang. & Literature	149	83.9	150	73.3	78.6
French	160	96.2	159	95.6	95.9
German	157	61.8			61.8
Home Economics Education	149	96.6	146	98.6	97.6
Industrial Arts Education	146	82.9	146	72.6	77.8
Mathematics	120	93.3	120	97.5	95.4
Media Specialist	144	97.9	144	92.4	95.2
Music Education	148	87.2	150	86.7	87.0
Physical Education	150	94.0	150	94.0	94.0
Social Studies	142	95.8	148	95.3	95.6
Spanish	159	89.3	151	87.4	88.4
Speech—Comm. & Theatre	145	71.7	143	73.4	72.6

^a The base number of questions, denoted by *n*, was the number classified with regard to content appropriateness by at least three judges. Questions classified by fewer than three judges were not included in the analysis.

^b The sum of the percentages for the two editions divided by 2.

^c Calculated by weighting the percentage for each component by the number of questions in the component as used in the NTE program, adding the products, and dividing by 270, the total number of questions in the Common Examinations.

^d Only one edition of German was included in the study.

The results of the knowledge estimation process can be summarized as follows:

1. A score of 548 on the Common Examinations was judged by the panel members to be the level that would be achieved by a minimally knowledgeable candidate.

2. Scores ranging from a low of 458 (Agriculture) to a high of 690 (Media Specialist) were judged by the panel members to be the level that would be achieved on the various Area Examinations by a minimally knowledgeable candidate. These scores are set out on p. 25 below.

3. The judgments obtained for test editions and half panels were generally consistent.

More generally, the study demonstrated that the content of nearly all the tests in the NTE was judged to be sufficiently congruent with the content of South Carolina teacher training programs to support the finding that the NTE are a fair measure of the knowledge sought to be imparted by those programs.⁴¹ Because the NTE are developed to national, rather than to local or state, specifications, some variation in the extent to which individual tests matched the content of specific teacher training programs in South Carolina was to be expected. The two area tests, in German and Speech-Communications and Theater, that ETS found to be "least related" to the curriculum in South Carolina, were singled out for further study before a determination could be made that they were appropriate for use in South Carolina.⁴² The 18 tests in other subjects included in the study and the Commons as a whole are all plainly appropriate for use in making certification decisions in South Carolina.

⁴¹ The REPORT characterized the relationship of specific tests to the teacher training curriculum in South Carolina as falling into one of four categories: very closely related (eight Area tests and two tests on the Commons), closely related (four Area tests and three tests on the Commons), reasonably related (six Area tests), and least related (two Area tests and one test on the Commons). REPORT, pp. 152, 157.

⁴² One expert, Dr. Guion, thought that this was a very conservative approach. He testified as follows:

Q: In your opinion, are the conclusions [stated in the REPORT] adequately supported by the data set out in the report?

A: Yes, they are I think a bit too conservative in their conclusions, and have made what I interpreted as rather disparaging comments about two of the tests [German and Speech] that I thought were not merited.

Guion deposition, p. 26 (Ex. S-5).

Moreover, the study demonstrated that, based on the judgments of the Knowledge Estimation Panels, a composite score representing the minimum level of academic knowledge that should be necessary for completing teacher training programs and teaching effectively in South Carolina can be set for each of the subject areas covered by the NTE. These composite scores of minimally qualified candidates are set forth in the REPORT, and range from 1006 for Agriculture to 1238 for Media Specialist—higher in all cases than the 975 score previously used by South Carolina for initial certification.

After setting forth the findings justified by the study, ETS emphasized that South Carolina's use of the validity study in establishing new certification requirements could properly reflect considerations not specifically addressed in the study. For example, ETS identified some aspects of the measurement process itself that could provide a rational basis for some departure by South Carolina from the composite scores emerging from the panel judgments. Most importantly, ETS discussed the relevance of the standard error of measurement for NTE composite scores and summarized the risks associated with the establishment of particular composite scores.⁴³ In addition, ETS cited other legitimate educational considerations, such as the demand for teachers in particular fields and the desirability of maintaining an adequate pool of minority group teachers, which could be weighed by South Carolina in establishing new certification requirements for teacher candidates in South Carolina.⁴⁴

F. Action by South Carolina State Board of Education

The State Board of Education officially reviewed the REPORT of the ETS validity study at its meeting of January 9, 1976. The REPORT was referred to the Board's Committee on Teacher Recruit-

⁴³ REPORT, pp. 121-37 (Ex. S-MAN-32B). The standard error of measurement is a statistical concept that refers to the range within which a test taker's score would likely fall on any particular testing occasion. A candidate's true score, as opposed to his reported score, is a hypothetical score that represents the mean score he would earn if he took an infinite number of equivalent forms of the test. The standard error of measurement defines the range around this true score within which his reported scores would fall. The REPORT sets forth the standard errors of measurement for NTE composite scores (Table 32. *Id.* at 123) and displays the probabilities of erroneously rejecting applicants with true scores greater than a specified value and accepting applicants with scores less than that value (Table 33, *Id.* at 124).

⁴⁴ *Id.* at 178-83.

ment, Training and Compensation and the Department staff for critical review and evaluation.

At the request of the Committee, the Department staff developed data regarding the racial characteristics of the persons certified and hired in particular subject areas and assessed the impact of possible cutoff scores.⁴⁵ The Department's measurement expert also reviewed the study from a technical standpoint and concluded that it provided a professionally acceptable basis for action by the Board.⁴⁶ After several meetings regarding the study, the Committee prepared its report for consideration by the full Board.

The recommendations of the Committee were adopted by the Board on June 11, 1976. The Board concluded generally that the ETS study provided an acceptable basis for decision making with respect to the minimum score requirement aspects of the State's certification system.⁴⁷ Having decided that the results of the study required changes in the certification system, particularly modification of the minimum score requirement of 975, the Board reviewed the various considerations relevant to the establishment of new minimum score requirements. Among these considerations were the standard error of measurement of the NTE, sampling error in the study, consistency of the results, supply and demand for teachers, racial composition of the teacher force, effect on teacher training programs, and administrative convenience. Of primary importance to the Board were the two risks associated with any particular score: the risk of certifying a candidate who does not have the minimum qualifications defined by the study and the risk of not certifying a candidate who does have these minimum qualifications.

The Board established new minimum composite scores as a requirement for certification in those areas for which an appropriate teaching area examination exists and has been validated for use in South Carolina. These scores, as well as those supported by

⁴⁵ Maynard deposition, Vol. V, pp. 46-65 (Ex. P-91). Some of this data is set out in Exhibits P-M-80C and P-M-80D.

⁴⁶ Report of the Committee on Teacher Recruitment, Training and Compensation (hereinafter "Committee Report") (Ex. S-28, S-32)

⁴⁷ In making this judgment the Board relied not only upon the experts used by ETS in preparing and reviewing the study but also upon an independent evaluation of the study by the Department's staff. Committee Report, p. 3. (Ex. S-32).

the study and the difference in standard error of measurement units between the two scores, are as follows:

Area	New Score	Study Score	SEM Units ⁴⁸
Agriculture	940	1006	-2
Art Education	1040	1104	-2
Biology and General Science	1115	1196	-3
Business Education	1071	1167	-3
Chemistry, Physics, and General Science	1100	1158	-2
Early Childhood Education	1052	1124	-2
Education in the Elementary School	1132	1194	-2
Education of the Mentally Retarded	1120	1234	-3
English Language and Literature	1041	1069	-1
French	1144	1210	-2
Home Economics Education	1075	1143	-2
Industrial Arts Education	1100	1162	-2
Mathematics	1113	1203	-3
Librarian-Media Specialist	1178	1238	-2
Music Education	1020	1086	-2
Physical Education	1125	1195	-2
Social Studies	1086	1148	-2
Spanish	1132	1163	-1

As is apparent, the Board established scores in every area one, two or three standard errors of measurement *lower* than those that were supported by the validity study.⁴⁹ These new minimum scores reflect the Board's assessment of the risks acceptable in each subject area in light of the subject involved, the available supply of teachers, and the desired racial composition of the teaching force. The Board's new scores are extremely conservative: reduction of the score supported by the study by two standard errors of measurement involves a risk of rejecting a maximum of two candidates out of 100 whose true scores are equal to the study score; reduction by three standard errors of measurement involves an even lesser risk (less than one in 1,000) of not certifying a minimally qualified applicant. The new certification requirements become effective November 1, 1976.⁵⁰

⁴⁸ The risks considered by the Board are defined by reference to the standard error of measurement. By definition, there are 68 chances in 100 that a candidate's observed score will be within one standard error of measurement of his theoretical true score; 95 chances in 100 that the observed score will be within two standard errors of the true score; and 99 chances in 100 that it will be within three standard errors of the true score.

⁴⁹ The Board established a minimum score of 510 on the Common Examinations for certification in the areas for which a teaching area examination either does not exist or has not been validated for use in South Carolina. Ex. S-28, S-32.

⁵⁰ Ex. S-32.

II. THE VALIDITY STUDY PROVIDES A RATIONAL BASIS FOR SOUTH CAROLINA'S NEW CERTIFICATION REQUIREMENTS THAT COMPLIES FULLY WITH APPLICABLE CONSTITUTIONAL AND STATUTORY STANDARDS

South Carolina's establishment of new certification requirements in reliance on the ETS validity study can be measured against two legal standards: the standard required by the Fourteenth Amendment to the United States Constitution and the one required under Title VII of the Civil Rights Act of 1964. Since the recent decision of the Supreme Court in *Washington v. Davis*, 44 U.S. L.W. 4789 (June 7, 1976), there can be no doubt that these standards are different and that the Constitutional requirements are less rigorous than the obligations imposed on certain test users by Title VII. It is clear that the Constitutional standard, requiring a rational basis for the challenged use of a test, applies to South Carolina's present use of the NTE; it seems equally clear that the ETS validity study and the other record evidence demonstrate that South Carolina's action fully complies with the Fourteenth Amendment. It is less clear whether Title VII applies to South Carolina's use of the NTE for certification purposes; if it does, however, the validity study provides ample evidence that South Carolina is not engaging in a discriminatory practice as defined by Title VII.

A. South Carolina's New Certification Rules Meet the Requirements of the Fourteenth Amendment

The Supreme Court held in *Washington v. Davis*, *supra*, that state laws or regulations cannot be invalidated under the Fourteenth Amendment solely because of their differential impact on persons of different racial groups; it is necessary rather that the State action be proven to have a discriminatory purpose. Assuming that no such purpose can be documented in this case,⁵¹ the Fourteenth Amendment requires only that South Carolina's use of the NTE be rationally related to the accomplishment of a legitimate governmental objective. The record evidence, particularly the ETS validity study and its use by South Carolina in revising its certification requirements, complies with this fundamental Constitutional requirement.

⁵¹ This issue is being fully briefed by the parties and ETS does not take a position as to its proper resolution. It is important to note, however, that plaintiffs in their Proposed Findings of Fact have not challenged the NTE as racially biased or maintained that the new certification requirements set by the Board on June 11, 1976 were motivated by a discriminatory intent.

1. *The Constitutional Standard Requires that South Carolina's Certification System Be Rationally Related to a Legitimate Government Objective.*

Although the courts may vary in the precise formulation of the standard, the fundamental test under the Fourteenth Amendment pertinent to this case cannot be seriously questioned. It is the standard applied by the three-judge court in the companion case of *United States v. North Carolina*, 400 F. Supp. 343 (E.D.N.C. 1975). The court there concluded that North Carolina's failure to validate its use of a cut-off score on the NTE violated the Fourteenth Amendment. According to the court, the

selection of a cut-off score that has not been shown to reveal teacher competency and that has a disparate impact on blacks '... is not reasonably related to any proper governmental objective,' and constitutes a burdensome and arbitrary denial of equal protection under the fourteenth amendment. *Id.* at 349.

This formulation of the Constitutional standard is defined further by the rulings in *Tyler v. Vickery*, 517 F.2d 1089 (5th Cir. 1975), *cert. denied*, 44 U.S.L.W. 3719 (June 14, 1976); and *Washington v. Davis*, 44 U.S.L.W. 4789 (June 7, 1976).

In *Tyler v. Vickery*, the Fifth Circuit rejected a challenge under the Fourteenth Amendment to actions of the Georgia Board of Bar Examiners in certifying candidates for admission to the bar in part on the basis of performance on written tests. The court affirmed the district court's use of a rational relationship test, holding both that the State was pursuing a legitimate objective and that appropriate written examinations could be used to achieve this objective. The court stated:

Appellants concede, as they must, that the state has a legitimate and substantial interest in excluding from the practice of law those persons who do not meet its standards of minimal competence and that the Georgia examination, as presently constituted, tests skills and knowledge which have a "logical, apparent relationship" to those necessary to the practice of law.

While appellants valiantly argue otherwise, these facts are sufficient in themselves to establish the rationality of the bar examination in the constitutional sense. If a state has the right to insist on a minimum standard of legal competence as a condition of licensure, it would seem to follow *a fortiori* that it may require a demonstration of such competence in an exami-

nation designed to test the fundamental ability to recognize and deal with legal principles. *Id.* at 1101.

No validity study was offered to support the particular minimum score adopted by the bar examiners; the court concluded that their "considered judgments" on this issue were sufficient to meet the Constitutional test. *Id.* at 1102. Accordingly, the court held that the Georgia bar examination satisfied the rational relationship standard under the Fourteenth Amendment.

A similar analysis is reflected in *Washington v. Davis*, the Supreme Court's most recent decision evaluating a test under the Constitutional standard.⁵² In concluding that the District of Columbia police department could properly use a verbal ability test as part of its recruit selection program, the Supreme Court emphasized that "the test is neutral on its face and rationally may be said to serve a purpose the government is constitutionally empowered to pursue." 44 U.S.L.W. at 4794. In addition, the court affirmed the district court's conclusion "that [the test] was directly related to the requirements of the police training program and that a positive relationship between the test and training course performance was sufficient to validate the former, wholly aside from its possible relationship to actual performance as a police officer." *Id.* at 4795.⁵³ In the absence of any discriminatory intent, the Supreme Court held that such a rational relationship was sufficient to comply with the Constitutional standard.

2. South Carolina's New Certification Requirements Have the Necessary Rational Basis.

Unlike the *North Carolina* case, the record developed in this proceeding provides ample evidence that South Carolina's new certification requirements meet the Constitutional test. The ETS validity study and its use by South Carolina demonstrate that the State's action is rationally related to the attainment of a legitimate

⁵² The case arose under the Fifth Amendment because it challenged action by District of Columbia and Federal officials. The standard under the Fifth Amendment is the same, however, as that under the Fourteenth Amendment for state action.

⁵³ The district court's reliance upon a validity study and other evidence to document a direct relationship between the test and the police training program was also accepted by the Court. 44 U.S.L.W. at 4795. There was no validity study, however, to support the minimum score requirement on the test which was left intact by the Court's opinion.

governmental objective—the assurance that only persons with the minimum amount of academic knowledge necessary to teach effectively are certified by South Carolina.

- a) *The validity study adequately documents the relationship between NTE scores and effective teaching under the relevant professional and legal standards.*

(i) *professional standards*: The ETS validity study complies with the pertinent professional standards contained in the APA STANDARDS and DIVISION 14 PRINCIPLES. These standards were developed under the auspices of the principal professional organizations in the field of testing and measurement to provide guidance to test users and others regarding the proper and rational uses of tests. Compliance with these standards confirms that the ETS validity study provides an acceptable foundation for South Carolina decision-making with respect to its certification program.

As discussed in Section I, the content validity design adopted by ETS was entirely appropriate under professional standards. This in the APA STANDARDS (at p. 26) and the DIVISION 14 PRINCIPLES (at p. 3). Nothing in these professional standards suggests that content validation was an inappropriate choice of validation method. Indeed, it was probably the only available validation design, given the purpose for which the NTE are designed and the objectives of the validity study. As the APA STANDARDS note, “[c]ontent validity is most commonly evaluated for tests of skill or knowledge,” that is, for “achievement testing.” APA STANDARDS at p. 28. Since the validity study was aimed at determining the propriety of using the NTE in evaluating graduates of South Carolina institutions, measuring the content of the NTE against the content of the teacher training programs in South Carolina was an indispensable component of the validity study.

Moreover, the design of the validity study was carried out in full accord with the professional standards. Evidence of content validity consists of showing that what a test measures constitutes “a representative sample of behavior to be exhibited in a desired performance domain.” APA STANDARDS at p. 28. In the ETS study the performance domain was specified as the content of the teacher training programs at 25 South Carolina institutions—a definition of performance domain fully consistent with the professional standards. The use of expert judgments to make the necessary comparisons of the

NTE and the teacher training programs is expressly permitted by the APA STANDARDS. Thus,

If a test is used to estimate achievement in American history in Grade 12, the performance domain is less objectively defined. Given agreement on instructional objectives, it could be defined in terms of the types and quantities of the skills, facts, and concepts of American history, as determined by the pooled judgments of authorities, experienced teachers, and competent curriculum makers in that field. APA STANDARDS at p. 28.

No real challenge can be advanced regarding the study's compliance with the relevant professional standards. The independent reviewers asked by ETS to critique the study design considered the study to be fully consistent with professional standards; one commented that the study was of "excellent technical quality,"⁵⁴ and another observed that the study "went beyond what would be called 'professionally acceptable.'"⁵⁵

Dr. Guion, one of the authors of both the APA STANDARDS and the DIVISION 14 PRINCIPLES, testified as follows:

Q Well, in your opinion does the study described in the report meet the relevant requirements set out in the APA Standards?

A Yes. And again it goes beyond them.

* * *

Q Dr. Guion, in your opinion does the study described in the report meet the relevant requirements set out in the Division 14 provisions?

A Yes.⁵⁶

(ii) *legal standards*: As *Tyler v. Vickery*, 517 F.2d 1089 (5th Cir. 1975), *cert. denied*, 44 U.S.L.W. 3719 (June 14, 1976), indicates, a validity study is not necessary to demonstrate that a particular test use is permitted under the Constitution. The validity

⁵⁴ Letter from Linn to Stewart, dated Sept. 27, 1975 (Ex. P-MAN-18).

⁵⁵ Guion deposition, p. 24 (Ex. S-5).

⁵⁶ Guion deposition, pp. 27-28 (Ex. S-5). For a summary of Dr. Guion's work in drafting the professional standards, see *Id.* at 16-23 and Dr. Guion's professional curriculum vitae (Ex. S-MAN-34).

study in this case, however, provides reliable and persuasive evidence that South Carolina's certification requirements easily meet the Constitutional test.⁵⁷

In *Tyler v. Vickery* the court pointed out, "[t]he hallmark of a rational classification is not merely that it differentiates but that it does so on a basis having a fair and substantial relationship to the purposes of the classification." *Id.* at 1099. More particularly, the court defined "two criteria of a rational examination" against which to review the Georgia bar examination; "a rationally supportable examination should (1) be designed for the purpose for which it is being used, and (2) utilize a cut off score related to the quality the examination purports to measure." *Id.* at 1101, 1102.

The validity study amply demonstrates a rational basis for South Carolina's continued use of the NTE. Indeed, the two types of panels utilized by ETS in the study produced and analyzed expert judgments responsive to the two criteria stressed by the Fifth Circuit. The Content Review Panels evaluated whether the NTE, designed to test academic preparation against a national norm, could properly be used to measure the academic knowledge of applicants trained at one of the 25 South Carolina teacher training institutions. The Knowledge Estimation Panels provided data as to the composite scores in various areas of specialization which would mark the minimally knowledgeable teacher candidate whom the State would be willing to have teach in South Carolina schools.

Compliance with standards such as those defined by the Fifth Circuit is all that is required to meet the rational relationship test. There is no requirement under the Fourteenth Amendment, for example, that a particular kind of validity study—or indeed any validity study—be conducted. The Supreme Court observed in *Washington v. Davis*, 44 U.S.L.W. 4789 n. 13 (June 7, 1976), that it "appears beyond doubt by now that there is no single method for appropriately validating employment tests for their relationship to job performance." And the fact that a particular study could be improved is irrelevant under the Fourteenth Amendment; "the focus of the rational relationship test is not whether the state has superior

⁵⁷ It is not necessary, therefore, to explore whether such a rational basis could be found in the absence of the ETS validity study. Nor is it required to differ with the Fifth Circuit's conclusion that under certain circumstances no validity study of any kind is necessary to meet the Constitutional test.

means available to accomplish its objectives, but whether the means it has chosen is a reasonable one." *Tyler v. Vickery, supra* at 1102.

Such a definition of the rational relationship test is particularly critical when the use of standardized tests by state authorities is being challenged. Properly used, standardized tests provide an objective measure that is open to inspection and easy to comprehend. From a standpoint of fairness, they are usually preferable to subjective measures because of their substantially higher reliability and potential for higher validity.⁵⁸ Therefore the legal standards applied to standardized tests should not be so stringent that the use of tests is made economically or otherwise impractical.

Beyond these concerns is the need to state legal standards in a way that does not impede progress within the measurement profession or prevent test users from using new validation techniques as they are developed. Validation techniques have been revised and refined over the years. Various theories have obtained currency in the profession and then declined in acceptance as more persuasive theories were developed. The measurement profession has continued to search for more accurate ways to define what tests measure and describe how they measure it. "Experience teaches that the preferred method of today may be the rejected one of tomorrow." *Vulcan Soc'y of N.Y. Fire Dept., Inc. v. C.S.C.*, 490 F.2d 387, 394 (2d Cir. 1973).

The validity study in this case represents one such new validation method and reflects the growing interest within the measurement profession in content validation.⁵⁹ Such new developments would be adversely affected if the legal standards implementing the rational relationship test are too specific or stated exclusively in terms of current validation techniques or theories. It is a sounder approach to follow *Washington v. Davis, supra*, and *Tyler v. Vickery, supra*, and define the legal standards generally in terms of the traditional requirement of rationality so that the necessary flexibility and room for experimentation can be preserved.

⁵⁸ Wilson, *A Second Look at Griggs v. Duke Power Company: Ruminations on Job Training, Discrimination, and the Role of the Federal Courts*, 58 VA. L. REV. 844, 873 (1972); Note, *Employment Testing: The Aftermath of Griggs v. Duke Power Company*, 72 COLUM. L. REV. 900, 924 (1972).

⁵⁹ *E.g.*, Guion deposition, pp. 37-38 (Ex. S-5).

- b) *The use of the validity study by the State Board of Education is rationally related to its objective of providing a sufficient supply of adequate teachers.*

The State's use of the validity study to establish new certification requirements provides further evidence that the rational relationship test of the Fourteenth Amendment has been abundantly satisfied. This is reflected not only in the Board of Education's evaluation of the validity study but also in its consideration of numerous other relevant factors.

Before acting upon the data and analyses contained in the ETS validity study, the Board of Education asked for a critical review of the study by one of its standing Committees and the staff of the State Department of Education. The first step in this process was a review with ETS representatives regarding the design, execution and conclusions of the study. With this additional explanation in hand, the staff conducted its own examination of the study. According to the report of the Board Committee, the review persuaded the professional Department staff that the ETS study was a professionally conducted validity study which could serve as a basis for decision making by the Board of Education.⁶⁰

The Board also considered a wide range of other factors related to its overall objective of ensuring that South Carolina schools would have an adequate supply of qualified teachers in the various areas of specialization. The Department staff developed data for the Committee relating to the supply and demand for teachers in particular areas, the possible racial impact of particular cut-off scores, and the risks associated with adjustments (either way) of the scores supported by the validity study. With this data before it, the Committee then debated what minimum scores to substitute for the earlier requirement of 975. The Committee's recommendations were in turn adopted by the Board on June 11, 1976.

The final action by the Board in setting new certification requirements reflects a reasoned evaluation by the Board of the validity study and the State's educational objectives. First, the Board established new minimum scores on the NTE generally within the framework set forth in the validity study. Second, the Board established different scores for different specialties, thereby reflecting the fact that identical scores on different NTE Area Examinations

⁶⁰ Exhibits S-28, S-32.

are not equated with each other. Third, the Board's action reflected individual consideration of each specialty for which the NTE was being validated, including review of the supply and racial characteristics of the teaching force available in each area. Fourth, the Board reduced all of the scores supported by the panel judgments in the study and thereby lessened the risk that an applicant with the requisite qualifications (as measured by the study) would be denied certification. Lastly, the Board's action was motivated by a desire to ensure that the teachers certified under the new system would have the minimum amount of academic preparation necessary to teach effectively. In sum, the Board's actions comply fully with the rational relationship test required by the Fourteenth Amendment.

B. The Validity Study Provides Adequate Evidence of Non-Discrimination Under Title VII

If South Carolina's certification requirements are sufficient under the Constitutional standard, they must then be examined under the different standards applied under Title VII of the Civil Rights Act of 1964. The parties differ on whether Title VII applies to the certification system used by South Carolina, specifically whether the State is to be treated as an employer (or employment agency) under Section 701(b)-(c) of the Act, 42 U.S.C. § 2000e(b)-(c). Even if it is determined that Title VII applies, the ETS validity study is persuasive evidence that the State's use of the NTE for certification purposes complies with Title VII's requirements.

1. Title VII Requires Proof by Defendants that Certification Decisions Are Made on a Basis Other than Race

The operative language of Title VII is straightforward: "It shall be an unlawful employment practice for an employer . . . to fail or refuse to hire . . . any individual . . . because of such individual's race" 42 U.S.C. § 2000e-2(a)(1). The Supreme Court has required in employment testing cases that the defendant bear the burden of proving nondiscrimination once sufficient evidence of differential impact by race has been offered by plaintiff. *Albemarle Paper Co. v. Moody*, 422 U.S. 405, 425 (1975). This burden requires the defendant to show that the challenged test (or other employment practice) is job-related, *i.e.*, it measures some skill or attribute required on the job and that the data displaying this relationship have statistical and practical significance.

Since South Carolina's certification system creates an adverse impact by race because a lower percentage of black applicants than white applicants are certified, the principal issue under Title VII (if it is applicable) is whether the State's reliance on the ETS validity study satisfies the pertinent Title VII standards. Plaintiffs urge that the standard applicable to the State's use of the NTE in reliance on this study is simple: total and inflexible compliance with the EEOC Guidelines. If the evidence offered by the defendant, most particularly the validity study, does not comply literally with each paragraph of the Guidelines, plaintiffs urge that this Court must find a violation of Title VII. There are several reasons why such exclusive reliance on the EEOC Guidelines is not warranted.

First, the Guidelines are out of date. They were first adopted by the EEOC on August 24, 1966, nearly a decade ago. *See* 35 Fed. Reg. 12333 (August 1, 1970). It was four years before the agency published the Guidelines on August 1, 1970, 29 C.F.R. Part 1607, and some revisions were made prior to publication. However, the basic approach reflected the professional views and technology available in 1965 when the Guidelines were first drafted. Those views and that level of technology have changed significantly over the past decade.⁶¹ There are extensive efforts underway to amend or replace the existing Guidelines and several drafts of new Guidelines, proposed by various federal agencies, are in the record.⁶² These

⁶¹ For example, the concept of differential validity was sponsored during that period by Dr. Guion, one of the experts deposed by defendants, and the Guidelines reflect his belief at that time that it represented an important new tool. Guion deposition, pp. 44-45 (Ex. S-5). Dr. Guion and most other measurement experts have changed their view about the utility of differential validity, because evidence developed in empirical studies indicated that the underlying hypothesis could not be proved. As Dr. Guion himself testified:

... the history of the research since I first proposed differential prediction in 1966 is that that was a beautiful hypothesis that doesn't work, the data simply do not fit the hypothesis, the evidence is very underwhelming. *Id.* at 43-45.

The Guidelines have not been revised and continue to demand evidence with respect to differential validity (*See* discussion *infra* at pp. 43-44).

⁶² For example, the EEOC has itself recently proposed a new draft of its own guidelines. Guion deposition, pp. 18-19 (Ex. S-5); Exhibit S-G-2. Similarly, the Equal Employment Opportunity Coordinating Council, a federal agency distinct from the EEOC, has circulated several drafts of proposed Uniform Guidelines on Employee Selection Procedures that would replace the several different sets of guidelines now maintained by various government agencies with overlapping jurisdiction. The most recent draft is dated September 24, 1975. Guion deposition, p. 21 (Ex. S-5); Exhibit S-G-3.

drafts reflect the lack of agreement, even within the federal regulatory community, as to the appropriateness of certain types of requirements.⁶³

Second, the Guidelines do not apply with equal specificity to the various methods of developing evidence available to an employer. Standards with respect to criterion-related validity, which was well known and widely used ten years ago, are detailed and specific; standards with respect to content validity are quite general; and there is virtually no comment on construct validity.

Third, the Guidelines are different, in important respects, from accepted professional standards in the field of testing and measurement. The APA STANDARDS and the DIVISION 14 PRINCIPLES were published to guide professionals in conducting validity studies under the same circumstances intended to be covered by the Guidelines. These present a more even-handed and specific treatment of the three principal types of designs for validity studies than the Guidelines, and are also substantively different. Further, the professional standards recognize the need for flexibility and reliance on expert opinion. They do not purport to apply to every case or to set out legal standards.⁶⁴

⁶³ For example, compare the EEOC's new draft "Guidelines on Employee Selection Procedures" (Ex. S-5, S-G-2) with the Coordinating Council's draft "Uniform Guidelines on Employee Selection Procedures" (Ex. S-G-3).

⁶⁴ The introduction to the APA STANDARDS goes out of its way to make this point:

This document is prepared as a technical guide for those within the sponsoring professions; it is *not* written as law. What is intended is a set of standards to be used in part for self-evaluation by test developers and test users. An evaluation of their competence does not rest on a literal satisfaction of every relevant provision of this document. The individual standards are statements of ideals or goals, some having priority over others. Instead, an evaluation of competence depends on the degree to which the intent of this document has been satisfied by the test developer or user. APA STANDARDS at 8 (Emphasis in original).

Similarly, the DIVISION 14 PRINCIPLES specify that they "are *not* intended to . . . freeze the field to prescribe practices, nor to limit creative endeavors." DIVISION 14 PRINCIPLES at 1 (Emphasis in original). They go on to provide that:

The last point deserves emphasis. Traditional technology calls for a showing that (a) assessments made by a particular method (or combination of methods) are useful for predicting behavior in some aspects of employment and (b) that the predictions can be made within an acceptable allowance for error (usually expressed in terms of percentage of misclassification or correlation coefficients). Principles presented here are stated in the context of the traditional approach The traditional approach is emphasized be-

Fourth, the Guidelines, if applied inflexibly, do not encourage or even permit innovation in this field. The study produced by ETS is a new design. It should be considered with respect to three essential elements: (1) whether the design of the study was professionally acceptable and adequate to the task; (2) whether the execution of the study design was such as to provide reliable results; and (3) whether the analysis of the data was performed according to professionally acceptable methods. Proof with respect to compliance with the Guidelines should be regarded as relevant to those issues. It should not be controlling.

2. *The ETS Validity Study Provides Adequate Evidence, Under the EEOC Guidelines, of a Non-discriminatory Basis for Certification Decisions*

Although compliance with the Guidelines should not be the sole determinant of legality under Title VII, the ETS validity study in fact satisfies the minimum standards set out in the Guidelines for validation. Those standards may be summarized as follows: (a) a criterion-related validity study must be done if that is feasible, § 1607.5(a); (b) a differential validity study must be done if that is feasible, § 1607.4(a)-(b); (c) a content validity study, if standing alone, must demonstrate that the tests consist of suitable samples of at least one essential element of satisfactory job performance, § 1607.5(a). Each of these standards has been met, as described more fully below.

a) *Requirement with respect to criterion-related study*

The Guidelines provide:

For the purpose of satisfying the requirements of this part, empirical evidence in support of a test's validity must be based on studies employing generally accepted procedures for determining criterion-related validity . . . [Other evidence of validity] may also be appropriate where criterion-related validity is not feasible. § 1607.5(a).

The *sine qua non* of a criterion-related study is an adequate criterion. A criterion is the measure of job performance against which the test results are compared to determine whether there is a statistically significant correlation. The correlation describes the

cause its principles have been established through a long history. *Other approaches may be equally good or even superior*, but it seems premature to try to articulate formally the principles of their use. DIVISION 14 PRINCIPLES at 1 (Emphasis added.)

extent to which it can be concluded that persons who have adequate test scores will also perform adequately on the job. Because defendants are certifying persons to become teachers, the criterion for any study useful to them must be a measure of effective teaching.

There is no known measure of effective teaching and for that reason a criterion-related validity study is not feasible. Even plaintiffs' experts were virtually unanimous on this point:

Dr. Thorndike's testimony:

The attempt to validate empirically the [NTE] against actual measures of effectiveness as a teacher are fraught with so many practical and technical difficulties as to make a really satisfactory study of this type nearly impossible.⁶⁵

Dr. Cohen's testimony:

Q. Dr. Cohen, there isn't any existing measure of competent teacher performance in the business, is there?

A. Not that I know of.

Q. And I believe I recall from our interview last night that you told me that measures of competent teacher performance, as they may now exist, are what you describe as a bag of snakes in the business?

A. Yes. I was referring when I used that colorful phrase to the problems of the researcher who is trying to develop such instruments. It is an extraordinarily difficult problem because of the underlying lack of agreement in the field as to standards.

Q. So that you will agree that to date no adequate measure of competent teacher performance exists?

A. I certainly will.

Q. Development of a measure of competent teaching is dependent on the technology of teaching being developed, is it not?

A. That was my opinion that I expressed last night, yes.^{65a}

Dr. Bartlett's testimony:

Q. Can you think of any published works in professional literature presenting evidence of the successful development and validation of measures of overall teaching effectiveness?

[Dr. Bartlett:] I think I have said that I can't identify specific articles in the field. I have read some at the college level, again that I think give some indication of it, and I can't come up with the exact references.

⁶⁵ Thorndike written statement regarding the ETS validity study submitted at deposition, p. 3 (Ex. P-THO-2).

^{65a} Cohen deposition, p. 64 (Ex. P-111).

Q. Do you know of any such measurements that have been generally accepted throughout the profession?

A. Well, there are a number of standardized measures for this, but I can't say that they have been generally accepted.
 . . . " ^{65b}

Dr. Medley testified:

Q. Dr. Medley, your study commented on the lack of any reasonably adequate criterion of teacher effectiveness that could be used to validate the NTE. Do you know of any such adequate criterion that has been developed since you published this study in 1974?

A. No. ⁶⁶

The experts who contributed to the ETS validity study also confirmed that no adequate criterion of teacher effectiveness currently exists. ⁶⁷

There do exist procedures (such as supervisory ratings) that *purport* to assess teacher effectiveness. And there are instruments and procedures that describe some of the things teachers *do*. Some of plaintiffs' experts, undeterred by the limitations of each of these, have in varying degrees advocated their use as a criterion. There is no basis in the record for asserting (as plaintiffs do) that any of these instruments or procedures is sufficiently developed so to permit an acceptable criterion-related study.

None of plaintiffs' experts seriously advocated the use of supervisory ratings of teachers as a criterion measure. Although two experts thought that such ratings *might be* developed into useful measures, ⁶⁸ none really disagreed with Dr. Thorndike's observation that "typically judgments of supervisors in a work situation are quite undependable and unreliable." ⁶⁹

^{65b} Bartlett deposition, p. 123 (Ex. P-109).

⁶⁶ Medley deposition, p. 107 (Ex. P-112).

⁶⁷ *E.g.* Linn deposition (Vol. II), pp. 5-6 (Ex. P-116).

⁶⁸ Bartlett deposition, pp. 123-26 (Ex. P-109); House deposition, pp. 106-07 (Ex. P-87).

⁶⁹ Thorndike deposition, p. 22 (Ex. P-110); Bartlett deposition, pp. 113-14 (Ex. P-109); House deposition, p. 105 (Ex. P-87). The principal basis for Dr. House's assertion that the study should have used criterion-related methods was that an article published by ETS personnel disclosed an absence of high correlation coefficients in other criterion-related studies involving the NTE. House deposition, p. 27 (Ex. P-87); T. Quirk, *et al.*, A CRITICAL REVIEW OF RESEARCH RELATED TO THE NATIONAL TEACHER EXAMINATIONS (1972) (Ex. P-MAN-8).

[Footnote continued on page 40]

"Classroom observations systems" and student ratings of teachers were two related methods that were suggested as possible criteria.⁷⁰ The former generally record expert observations of what goes on in a class. The latter systematically record the reactions of students to teachers. There are perhaps several hundred or more different classroom observations systems⁷¹ and there are probably as many student ratings forms. While each may provide some information about what *goes on* in a classroom, or how students *react* to it, the record is barren of probative evidence that any of these systems measure teaching effectiveness or success.

With respect to student rating forms, Dr. Bartlett had to concede that a "student's description of a teacher doesn't necessarily tell you whether" what the teacher does "is good or bad."⁷² And, although he maintained that such reactions had been validated against teaching effectiveness, he could cite no professional studies to support his position.⁷³

Similarly, Dr. House would not say, except in the most circumspect way, that classroom observations systems measured teacher

⁶⁹ [Continued]

This is somewhat inconsistent with Dr. House's recognition that supervisory ratings are not generally useful measures of teacher effectiveness. As Dr. House later acknowledged, most of the studies discussed in Exhibit P-MAN-8 used supervisory ratings as a criterion measure. In view of the unreliability of such ratings, the validities reported in Exhibit P-MAN-8 might very well be misleadingly low. Dr. House acknowledged as much. House deposition, pp. 105-07 (Ex. P-87).

⁷⁰ *E.g.*, House deposition, pp. 48-49; 108-13 (Ex. P-87).

⁷¹ *Id.* at 113.

⁷² Bartlett deposition, p. 120 (Ex. P-109).

⁷³ Dr. Bartlett was repeatedly pressed to identify the published studies on which his position was based. *E.g.*, Bartlett deposition 109-10, 117-18, 121-23 (Ex. P-109). Apart from his identification of "a methodological article in general" (*id.* at 108), an article dealing generally with "residual gain measures" (*id.* at 111) and a study dealing with "several classifications, I believe in government jobs, one of which was cartographers" (*id.* at 117), all Dr. Bartlett could come up with was an annotated bibliography whose name he "forgot to take down" (*id.* at 109) and a doctoral dissertation at the University of Maryland (where Dr. Bartlett teaches) that has not yet been completed (*id.* at 110).

While his description of the unfinished doctoral dissertation was somewhat glowing, he concluded that it, too, dealt only with the evaluation of "college teachers" by college students (*id.* at 119), that it had not yet been defended (*id.* at 110), and had not been circulated or subjected to criticism outside the Psychology Department of the University of Maryland. Thus, it has not been scrutinized by the profession in general and is unavailable for cross-examination with respect to its relevance to this case.

effectiveness.⁷⁴ Such systems, by their nature, do not even indicate whether students are in fact learning anything or whether what they are learning is correct. While they "tell what a teacher does, . . . there are varied opinions as to whether or not those behaviors are effective."⁷⁵

Dr. Thorndike suggested that, given resources in the range of "six figures," one might conduct a validity study using as criteria a "composite of the opinions of both supervisors and students." However, Dr. Thorndike clearly indicated that there were limits on the extent to which that composite would constitute a measure of effective teaching:

I would be reluctant to use either one without the other because I think either one sees only a partial picture and the two taken together would seem to me to be much more acceptable as indicating *at least how the teacher was perceived* in his or her functioning by both the persons who were responsible for that teacher and the persons who were effected by that teacher.⁷⁶

In view of the lack of agreement on what classroom observation systems and student rating forms *mean* in terms of teaching effectiveness, they are unsuitable as criteria for validating the NTE as a *measure* of teacher effectiveness. Indeed, a study based on either would be inconsistent with professional standards. As the APA STANDARDS make clear:

Criterion-related validity assumes that the criterion possesses validity. All too often, tests are validated against any available criterion with no corresponding investigation of the criterion itself. The merit of a criterion-related validity study depends

⁷⁴ Dr. House was somewhat obscure on this point:

Q. Well, are you then saying that . . . you don't know whether classroom observation is a measure of teacher effectiveness?

A. It can be under certain circumstances.

I mean, there are—there is no one teacher effectiveness. There are many different kinds of teacher effectiveness in classroom observations schemes. You *may* measure teacher effectiveness in a *certain sense of that term*.

Q. Then, the converse would be true that it may not measure teacher effectiveness if it may measure only in certain circumstances?

A. Well, it depends on—again, *it depends on* what the measure is, *what you are calling teacher effectiveness*. House deposition, pp. 110-11 (Ex. P-87) (Emphasis added).

⁷⁵ Linn deposition (Vol. II), p. 42 (Ex. P-116).

⁷⁶ Thorndike deposition, pp. 25-26. (Ex. P-110) (Emphasis added).

on the appropriateness and quality of the criterion measure chosen Criterion-related validity studies based on the 'criterion at hand,' chosen more for availability than for a place in a carefully reasoned hypothesis, are to be deplored. APA STANDARDS at 27.

The DIVISION 14 PRINCIPLES also underscore that "[c]riteria should be chosen, not on the basis of availability, but on the basis of importance and relevance." Guideline (A) (1) at 3. The record leaves little room for doubt that the use of student reactions and classroom observation systems, like supervisory ratings, are at least for the present simply "criteria at hand." None has been demonstrated to measure teaching success, and none is suitable for use in the type of study advocated by plaintiffs.⁷⁷

The record in this case, in short, does not in any respect demonstrate the existence of an accepted measure of teaching effectiveness or success suitable for use in a criterion-related study. In that case, there is nothing to be gained by enforcing a requirement that a criterion-related validity study be done. Professional standards are unequivocal on this point. The DIVISION 14 PRINCIPLES stress that:

It is not always possible to do a well-designed or even reasonably competent study; a poor study is *not* better than none at all. Several questions must be considered in deciding whether a study of criterion-related validity is feasible (*cf.* Stds. pp. 26-28).

* * *

. . . a reasonably valid, uncontaminated, and reliable criterion is assumed in criterion-related validation. If such a criterion measure cannot be developed, criterion-related validation is not feasible. Principle (C) (1) at 4-5 (Emphasis in original).

⁷⁷ Another possibility suggested by the record is the use of what are known as "residual gain measures" of student learning. To do a satisfactory study of this type, however, would require the investigator to attempt to isolate that portion of a student's learning gains over time that is attributable to the teacher (or teachers) to whom the student is exposed. This requires the elimination of those aspects of the student's measured performance that are attributable to prior learning or general ability and other factors that may affect learning such as school environment. Needless to say, a study of this sort, if it could be done, would be extraordinarily complex. And it could only identify and account for gains in those areas that could be measured by standardized tests. Thus, Dr. Thorndike, who estimated that to conduct a study using residual gain measures of a criterion "you would certainly need an amount that would run to six figures and, very possibly, to seven" (Thorndike deposition, p. 25), thought that the study would be "only partially" satisfactory. *Id.* at 24.

Undeterred by the present state of the art, plaintiffs urge that, even if this court finds that an adequate measure of teaching effectiveness does not now exist, one *could be developed*. It is probably true that some day someone will develop an adequate criterion. It is also true that large numbers of researchers, including many at ETS, have been working on that problem for a long time without success.⁷⁸ Given large amounts of time and money, perhaps this criterion could even if this court finds that an adequate measure of teaching effectiveness be developed relatively sooner rather than later.

However, the "feasibility" standard, even under the Guidelines, does not look to the long-term future. It must look to what is available or readily adaptable at the time the study is proposed. It is undisputed on this record that such a criterion does not now exist and that, accordingly, the Guidelines did not require ETS to conduct a criterion-related study.

b) *Requirement with respect to differential validity study*

The Guidelines provide:

[W]here technically feasible, a test should be validated for each minority group with which it is used. . . . § 1607.4(a).

"Differential validity" applies only in the context of a criterion-related validity study. As Dr. Manning testified:

Differential prediction refers to whether a test predicts unequally well rather than equally well or equally poorly for various sub-groups. It is usually used to refer to whether the validity of a test for one ethnic or racial group is comparable for the validity that is obtained for another such group. Differential prediction as a concept is relevant only to criterion-related validity studies. It does not have applications in instances of construct validity or in instances of differential, in content or in construct validity.⁷⁹

Dr. Manning was cross-examined by plaintiffs on this point and explained:

Q. May a differential validation study be conducted using the content validation method?

⁷⁸ The ETS research program is described at Manning deposition, written questions on re-direct examination, Q. 44 (S-8).

⁷⁹ Manning deposition, p. 206 (Ex. S-8). Dr. Guion also testified that "differential prediction" is not applicable to the ETS study design. Guion deposition, pp. 45-46 (Ex. S-5).

A. I don't think that is an applicable concept in the instance of content validation.

Q. What is the basis for your view?

A. One is not dealing with making predictions of a kind that one is talking about [when] using differential validation or prediction. One is dealing, in content validation, with the degree to which performances on the test in question are a representative sample of a domain of performance from which the test has been drawn and so, in particular, one starts with the definition of a domain and then deals with the process of establishing the representativeness of the materials on the test, that is the performance that is required by the test, showing that in the test is a representative sample of that domain. So we're not dealing with just an empirical prediction model of the kind that we're talking about when we talk about differential prediction or differential validation.

* * *

Q. Is it true that the content validation method could be conducted separately for different ethnic or racial subgroups with respect to a particular test?

A. No.⁸⁰

As demonstrated above, a criterion-related validity study is not feasible with respect to the job of teaching because there is no adequate criterion of success as a teacher. Therefore, differential validity is also not feasible.

c) Requirements with respect to content validity study

The Guidelines were drafted at a time when criterion-related validity work was well-known within the profession and a principal tool of industrial psychologists. The structure of the Guidelines reflects an emphasis on that method. Section 1607.5 is entitled "Minimum Standards for Validation" and has two subsections devoted to validity studies. Subsection (a) identifies criterion-related, content and construct validity as the alternatives available to employers in validating tests. With respect to a content validity study, it provides:

Evidence of content validity *alone* may be acceptable for well-developed tests that consist of suitable samples of the essential knowledge, skills or behaviors composing the job in question.
§ 1607.5(a) (Emphasis added).

Subsection (b) then sets out "minimum standards, as applicable, [that] must be met in the research approach and in the presentation

⁸⁰ Manning deposition, pp. 522-23 (Ex. S-8).

of results which constitute evidence of validity." These minimum standards are contained in five numbered parts of Subsection (b), each of which applies only to criterion-related validity studies.⁸¹

Dr. Guion, an expert deposed by plaintiffs, has extensive familiarity with the EEOC Guidelines and their intended application from his wide-ranging professional experience and participation in the drafting effort in 1970 with respect to the Guidelines. He testified as follows:

Q What provisions of the current E.E.O.C. guidelines are relevant to a study such as [the ETS validity study] report?

A One sentence that says that under certain circumstances evidence of content—I'm not sure if this is a direct quotation but it is pretty close—Under certain circumstances evidence of content validity itself may be acceptable. That is the only thing in that.

Q In your opinion does the study described in the report meet the relevant requirements in the current E.E.O.C. guidelines?

A There really aren't any requirements relevant to this in the E.E.O.C. guidelines. There's only that one sentence that deals with content validity, and the requirements for content validity are not spelled out in the current, the existing regulations.⁸²

⁸¹ Section (b) (1) discusses the sample selection for predictive studies (criterion data collected later) and concurrent studies (tests administered to present employees); Section (b) (2) requires measures to ensure that the test scores do not contaminate the criterion measures; Section (b) (3) deals with the criteria of employee adequacy and includes specific requirements with respect to rating forms; Section (b) (4) also deals with supervisory rating forms; and Section (b) (5) deals with differential validity.

⁸² Guion deposition, p. 29 (Ex. S-5). As Dr. Guion recently observed:

with a committee of my peers, I had a guilty hand in formulating the document that became, after mutation, the EEOC Guidelines on Personnel Selection Procedures. . . . Quite casually, the Committee inserted two sentences on content validity in that document. One said that content validity might be a permissible means of evaluating employment tests where criterion-related validation was not feasible. The other said that content validity might be a permissible means of evaluating employment tests even where criterion-related validation *might* be feasible. In retrospect, that seems not very enlightening! Guion, *Content Validity, the Source of My Discontent*—APPLIED PSYCHOLOGICAL MEASUREMENT—(to be published Fall, 1976) (Ex. P-G-1 at p. 1).

The ETS validity study meets the one-sentence requirement set out in the EEOC Guidelines with respect to content validity.

The NTE are "well-developed tests." They have been given several times a year in each year since 1940. Since 1963 they have been revised and up-dated continuously by means of a detailed, careful, professional process. Over 150 independent experts have contributed to the design of specifications or the formulation of questions in the past five years alone. In addition, 12 ETS testing and subject matter experts are assigned to the NTE program and these experts are provided support by ETS's extensive research effort on all aspects of testing.⁸³

The study demonstrates that the NTE also "consist of suitable samples of the essential knowledge . . . comprising the job in question." There is no dispute among the parties as to two of the three elements of this requirement. Ultimately, the job in question is the job of teaching in South Carolina public elementary and secondary schools. However, since South Carolina requires completion of an approved program, a requirement not challenged by plaintiffs, there is also involved here the "job" of completing that training program.⁸⁴ The "essential knowledge" for either of these jobs is knowledge of the subject matter measured by the NTE; psychological and sociological foundations of education, methods and principles of teaching, and the specific subject to be taught (art, reading, etc.). Plaintiffs make no argument that this knowledge is not "essential" to effective teaching.

A "suitable sample" under the Guidelines looks to the relative weight or emphasis in the test with respect to each of the subject matter areas sought to be tested. The plaintiffs contend that the NTE do not provide such a sample. This requirement of the Guidelines was addressed directly by the ETS study in not one but three

⁸³ The process by which the Common Examinations are developed is described in detail in Appendix A. The members of the Committees of Examiners for various tests in the NTE Program are identified in Exhibit S-MAN-4. The outside experts who have written questions during this period are listed in Exhibit S-MAN-7. The qualifications of the ETS staff experts assigned to the NTE Program are set out in Exhibits S-MAN-6 and S-MAN-40.

⁸⁴ As Dr. Guion testified:

In the present case, the only way in which a job analysis concept makes any sense is if you refer to the job of being a student in a teacher training program. Otherwise, it has no relevance. Guion deposition, p. 119 (Ex. S-5).

different measures. The judges were asked: (1) whether the questions in the NTE covered subject matter that 90% of the students at their institutions had an opportunity to learn; (2) whether the emphasis on topics in the curriculum at their institution was "more," "less" or "same" as the emphasis in the NTE; and (3) whether the test failed to cover any major topics that the curriculum covered. The results of the study confirm dramatically the suitability of the sample—86% of the questions on the Commons and over 90% of the questions on 13 of the 20 Area Examinations were found content appropriate.

d) *Plaintiffs' validity study*

On the issue of criterion-related validity versus content validity, there is in the record a series of exhibits with respect to a validity study performed by plaintiffs, admitted into evidence by stipulation of the parties, that demonstrates in a practical sense each of the arguments set out above.

Plaintiffs have known from the outset of this litigation that the feasibility of a criterion-related validity study would be a major issue. It was one of their principal contentions in the *North Carolina* case and has been raised by them in numerous other cases involving the NTE. They had a tactical choice: to wait on the sidelines and then seek to elicit expert testimony as to what might have been done, knowing that other experts would claim it could not have been done; or to demonstrate that a criterion-related validity study could in fact be done by doing a study, and at the same time hopefully proving another principal contention, that the NTE do not correlate with measures of teacher effectiveness.⁸⁵ Plaintiffs decided to do both. They marshalled their experts to offer extensive opinions on what hypothetically might be done; they also did a validity study.

In doing their study, plaintiffs faced all of the problems that ETS would have faced had it chosen to use a criterion-related design. The first is the choice of criterion. Plaintiffs had ample time to find or devise a criterion: one of the plaintiffs had determined by October 8, 1974 that a violation of Title VII existed;⁸⁶ another

⁸⁵ Plaintiffs have conceded that they have the capability to do such a study. See Answers to the Second Interrogatories of Defendant South Carolina to Plaintiff United States, filed April 5, 1976, Interrogatories Nos. 24-25, at 16-17.

⁸⁶ SCEA filed a Charge of Discrimination with the EEOC on October 8, 1974.

had commenced its investigation by October 19, 1974.⁸⁷ From October 1974 until May 1976 plaintiffs considered their alternatives, basically to use an existing criterion or to attempt to develop a new one. Their choice of the former was undoubtedly foreordained by the fact that developing a new criterion would be a very time-consuming and problematical task, and could not be accomplished in the two-year period which plaintiffs had to accumulate their evidence. Plaintiffs then reviewed the alternatives with respect to existing criteria and selected supervisory ratings as the most promising. In many South Carolina school districts standard rating forms are used by principals to assess teacher characteristics on a regular basis. Plaintiffs presumably structured their action to include the school districts as defendants, *inter alia*, in order to have access to the necessary records to complete their validity study, which were demanded and produced in due course.

As elaborated in Appendix B to this brief, the results of the plaintiffs' study are meaningless and useless. The study was such a complete failure that plaintiffs make no mention of it in their findings of fact, despite an extensive treatment of their contention that a criterion-related study is feasible. If ETS had attempted to perform such a study, plaintiffs' criticisms would have been long, loud and legion. Their own failure, however, has in no way diminished their vigorous insistence that ETS should have engaged in a similarly futile effort to conduct a criterion-related validity study where none was in fact feasible.

3. *South Carolina's Use of the Validity Study Was Proper Under the Guidelines*

The EEOC Guidelines contain several provisions with respect to the use of a validity study. Section 1607.5(c) provides:

In assessing the utility of a test, the following considerations will be applicable:

- (1) The relationship between the test and at least one relevant criterion must be statistically significant. . . .
- (2) In addition to established significance, the relationship between the test and criterion should have practical significance.

⁸⁷ The United States advised South Carolina by letter to Dr. Busbee dated October 19, 1974, that it was opening an investigation.

By their very terms, those requirements are directly relevant only to a criterion-related validity study.⁸⁸ To the extent that the concept of practical significance applies to the ETS validity study, South Carolina met that requirement with respect to the results of the content review aspect of the study.⁸⁹ Furthermore, several measures of the consistency of the results were devised, and each of these indicated that the results had a high degree of utility.⁹⁰

There are no readily available alternatives to the NTE. Grade point average is not a sufficient measure because of the variation in standards among the 25 teacher training institutions. Grades in practice teaching assignments suffer from numerous recognized deficiencies: they may be based on work other than in the classroom; there is a variation among supervisors in grading student teachers and different standards among different institutions; and the spread of grades in student teaching courses is very limited—almost all students receive a grade of A or B—so that the measurement of achievement of the requisite level of knowledge may not be accurate.⁹¹

South Carolina's care in the use of the validity study goes far beyond the requirements of the Guidelines. The Board considered not only the statistical characteristics of the study results and the possible alternatives, but the standard error of measurement of the NTE, the supply of and demand for teachers within the state, the racial composition of the teacher force, and the effect of the new certification standards on teacher training programs. Taking those factors into consideration, the Board lowered the score for every field below those recommended by the study by two or three standard errors of measurement in nearly every case.⁹²

⁸⁸ Both subsections refer to a relationship to a criterion, which would be present only in a criterion-related study.

⁸⁹ See discussion at pp. 20-21 *supra*.

⁹⁰ REPORT, pp. 92-101, 107-20 (Ex. S-MAN-32B).

⁹¹ Mahaffey, *Student Teaching Grades: Assessment and Use*, SCHOOL AND SOCIETY 5 (October, 1974); T. Quirk et al., A CRITICAL REVIEW OF RESEARCH RELATED TO THE NATIONAL TEACHER EXAMINATIONS 33 (1972) (Ex. P-MAN-8).

⁹² The scores recommended by the study and those adopted by the State are compared at p. 25 *supra*. The scores were lowered only one standard error for the teaching fields of English and Spanish.

Plaintiffs have devised a new test of utility which they claim renders South Carolina's reliance on the study improper, even if it were properly designed and conducted. Plaintiffs claim that tables of bivariate normal distribution should be used to test whether the use of a validity study is proper,⁹³ and they include several tables in their Proposed Findings of Fact purporting to show that selection of candidates using the NTE is no better than random selection would be. That claim is just another attempt to find some statistical measurement by which the study can be made to appear, on first impression, to have been improper.

A brief review of the merits exposes the critical defects in plaintiffs' analysis: (1) the use of bivariate normal tables is not required by the Guidelines, to which plaintiffs cling with such tenacity in other respects; (2) neither of the assumptions used in plaintiffs' application of the bivariate normal tables is realistic or has support in the record; (3) even these tables, given a complete interpretation (which was not provided by plaintiffs), support South Carolina's use of the study results; and (4) by using the risk tables provided in the study, South Carolina has given generous consideration to the problem of possible exclusion of qualified candidates.

The bivariate normal tables that plaintiffs claim must be used to assess utility have been used by psychometricians since 1939 when they were first published.⁹⁴ They were available for use when the Guidelines were first adopted in 1966. No mention is made of these tables or any requirement with respect to their use in the Guidelines. In no other case has the plaintiff United States urged that these tables must be used to establish the requisite utility under the Guidelines. These tables were brought to plaintiffs' attention by Dr. Medley who was the only expert to testify with respect to their use, and the possible contention with respect to these tables was not revealed by plaintiffs until May 14, 1976, the date of the Medley deposition, which was long after plaintiffs responded to defendants' contention interrogatories and long after the date for the close of discovery. This is a last-ditch effort by plaintiffs to find some statistical tool that will give credibility to their implacable opposition to any validity study with respect to the NTE.

⁹³ Plaintiffs Proposed Findings of Fact, ¶¶ 256-60.

⁹⁴ Taylor, et al., *The Relationship of Validity Coefficients to the Practical Effectiveness of Tests in Selection, Discussion and Tables*, 23 JOURNAL OF APPLIED PSYCHOLOGY 565-78 (1939) (Ex. P-MED-15).

Second, even if an application of the tables were a requirement under the EEOC Guidelines, plaintiffs have flagrantly misused them. The tables require two facts: the predictive validity coefficient of the test and a measure of the number of candidates who are in fact incompetent. Neither of these facts is available. In the absence of any criterion of effective teaching, no predictive validity study can be done and there is no predictive validity coefficient for the test.⁹⁵ Further, plaintiffs' own argument is that there is no adequate measure of the number of candidates who are in fact incompetent. In the past, the State has relied on a cutoff score of 975 and under those conditions about 20 percent of the candidates have been disqualified. However, as revealed by the results of the ETS validity study, the collective judgment of 456 South Carolina educators is that the 975 cutoff score was far too low, and under that assessment the 20 percent figure is also far too low because many incompetents were in fact certified.

Third, plaintiffs have chosen not to comment on the evidence from these tables, using their own assumptions, that indicates a high degree of utility in South Carolina's use of the NTE. Considering the reverse of the proposition that plaintiffs put to the tables (that is, what proportion of the candidates *not* selected were *actually not* qualified), the tables show the very significant result of 84 percent. Dr. Medley testified that, using the NTE under the assumptions of .50 validity and 10 percent of the candidates are actually incompetent, the use of the NTE will weed out a group of candidates 84 percent of whom *are actually incompetent*.⁹⁶ As the percentage of actual incompetents rises above 10%, the effectiveness of the NTE in screening out incompetents also rises. Given the interest of South Carolina in keeping incompetent teachers out of the classrooms, this result is significant indeed.

Fourth, South Carolina actually considered carefully and made generous allowance for the risk factor identified by plaintiffs. The REPORT included an extensive discussion of the risk of rejecting

⁹⁵ Dr. Medley assumed a predictive validity coefficient of .50 which was the highest actual coefficient achieved in previous predictive validity studies. That assumption has no basis whatsoever if previous studies were done with inadequate criteria. A defective criterion can cause a relatively low correlation when the true correlation would be quite high. Dr. Medley conceded that point. Medley deposition, p. 115 (P-112).

⁹⁶ *Id.* at 127-28.

qualified candidates,⁹⁷ defined by the standard error of measurement. As discussed above, however, the State has adopted a very general allowance for this risk by reducing the score levels recommended by the faculty members *two* standard errors of measurement for 12 of the tests, and *three* standard errors of measurement for four of the tests.⁹⁸ Two of the score levels were reduced by only one standard error. For the score levels set at two standard errors below the recommended level there are only two chances in 100 of failing to certify an applicant whose true score is equal to the score level recommended, only one chance in 100 of failing to certify a candidate whose true score is one half a standard error above the score level recommended; and one chance in a thousand or less of failing to certify a candidate whose true score is one or more standard errors above the recommended score.⁹⁹ For the score levels set at three standard errors below the recommended score level, there is one chance in 1000 or less of failing to certify a candidate whose true score is equal to or higher than the score level recommended.

Plaintiffs have failed to demonstrate that the validity study or South Carolina's use of it falls short of either the Constitutional or the Title VII standards. If the court so concludes, further analysis of plaintiffs' numerous criticisms of the validity study is unnecessary. In the next section of the brief, however, these criticisms are reviewed in detail so that the court may be fully satisfied that the ETS study has yielded reliable data upon which South Carolina certifying authorities could properly rely.

III. THE CRITICISMS OFFERED BY PLAINTIFFS' EXPERTS DO NOT PROVIDE ANY GROUND FOR DISQUALIFYING THE VALIDITY STUDY AS PROVIDING A RATIONAL BASIS FOR CERTIFICATION DECISIONS

A large portion of the voluminous record in this proceeding reflects plaintiffs' efforts to discredit the ETS validity study relied upon by South Carolina. The plaintiffs examined and cross-examined some eight expert witnesses in an effort to develop sufficiently numerous and persuasive criticisms of the validity study so that this court would conclude that the study could not be relied upon.

⁹⁷ REPORT, pp. 121-25 (Ex. S-MAN-32B).

⁹⁸ See p. 25 *supra* for identification of the tests.

⁹⁹ REPORT, Table 33, p. 124 (Ex. S-MAN-32B).

Much of the expert testimony elicited by plaintiffs is conflicting; it tends to prove what already is known: that every expert would have tinkered with the validity study in some respect and that no such study is beyond criticism. In addition, plaintiffs deposed 81 of the 456 South Carolina educators who participated in the study regarding its execution. None of this evidence provides any persuasive rationale for discrediting the validity study and indeed, supports the fairness and objectivity with which the study was done. In this section of the brief ETS summarizes the criticisms directed at the study's design, execution and data analysis, evaluates the record with respect to such criticisms, and demonstrates that the study represents a reasonable exercise of professional judgment by ETS that has produced reliable data upon which South Carolina authorities could responsibly establish their certification requirements.

A. The Design of the Validity Study Was Responsive to the Legitimate Objectives of South Carolina and Represented a Reasonable Exercise of Professional Judgment

Several of the criticisms on which plaintiffs rely most heavily are directed at the basic design of the ETS validity study. As discussed in Section II, plaintiffs' fundamental complaint that ETS and South Carolina should have used a criterion-related study ignores the record evidence that such a study is simply not feasible. Even assuming that a content validity approach is permitted, however, plaintiffs contend that the particular design utilized by ETS is seriously defective. They challenge, for example, the study's basic assumption that a single statewide standard of minimal competence is appropriate for graduates from all 25 teacher training institutions in South Carolina, the study's focus on the content of such teacher training programs rather than the job of teaching, the use as panel members of teacher educators rather than practicing teachers, the use of a majority-rule standard in determining the appropriateness of individual NTE questions, and the failure of ETS to conduct an appropriate pilot study. None of these criticisms provides any basis for concluding that the study design was other than a competent and thoughtful effort to develop reliable data responsive to the legitimate needs of South Carolina certification authorities.

1. The Validity Study's Use of a State-wide Standard of Minimal Competence Was Both Reasonable and Necessary

Although their argument is difficult to comprehend, plaintiffs appear to maintain that no statewide standard measuring academic

preparation can properly be imposed by South Carolina authorities.¹⁰⁰ To the extent that the judgments of the Knowledge Estimation Panels are directed at supporting such a standard, therefore, plaintiffs contend that the study design was fatally flawed. Plaintiffs seem to be arguing that, in light of the acknowledged variations among the 25 teacher training institutions in the State, the certifying authorities can inquire only whether a candidate for certification acquired the minimal amount of knowledge that he should have acquired at the particular institution which he attended.¹⁰¹ On the assumption, moreover, that predominantly black institutions differ as a group from predominantly white institutions in the content of their programs, plaintiffs contend that ETS erred in not collecting data to support this assumption. These contentions ignore the legitimate interest of South Carolina in evaluating *all* teacher candidates against a single objective standard of academic preparation and fail to comprehend that the NTE can properly be used to accomplish this objective.

- a) *South Carolina authorities have a legitimate interest in evaluating the academic preparation of graduates from different teacher training programs against a single statewide standard*

Although conceding the considerable variation among the 25 institutions currently preparing prospective teachers in South Carolina,¹⁰² plaintiffs fail to acknowledge the practical consequences of such variations for certification authorities charged with the responsibility of assuring that only those candidates with the requisite competence are certified to teach in the public schools. Instead, plaintiffs contend that, once having approved the curricula of such varied institutions, the State must necessarily certify every graduate who acquires the minimal amount of knowledge sought to be imparted by his school. Such a position ignores the pertinent facts displayed on this record and would dilute the State's certification process into a meaningless exercise that would yield no relevant information regarding teacher candidates which is not already reflected in their grade point averages.

Notwithstanding plaintiffs' assertions, South Carolina cannot be blind to the considerable variations that in fact exist among its

¹⁰⁰ Plaintiffs' Proposed Findings of Fact ¶¶ 219-26.

¹⁰¹ *Id.* at ¶ 217.

¹⁰² *Id.* at ¶ 223.

institutions. Certainly the State's failure to declare any of its 25 programs unacceptable under NASDTEC standards cannot require such an absurd result. In fact, the State is constantly evaluating its programs against several different sets of standards, reflecting the State's commendable interest in improving the quality of its teacher training. Four teacher training institutions in South Carolina have not yet been approved under NASDTEC standards; and only five institutions are approved under the more rigorous NCATE standards.¹⁰³ Plaintiffs' position is basically at odds with this effort by South Carolina to improve the quality of teacher training through the adoption of increasingly rigorous approval standards. Under plaintiffs' theory the graduates of any institution permitted to function in South Carolina can be evaluated only against the program at the institution, regardless of the approval standards which the institution does or does not meet.

Even assuming that all South Carolina teacher training programs are deemed acceptable by the State, plaintiffs fail to acknowledge the limited significance of such approval so far as it relates to the quality of the education offered in these programs. Notwithstanding their compliance with particular approval standards, there are wide variations among the standards for admission to and graduation from the 25 institutions.¹⁰⁴ Some have virtually open admissions policies; the quality of instruction available in each curriculum area differs from institution to institution, as do the standards of acceptable completion of a particular subject area. If plaintiffs' contentions were accepted, graduates of the more rigorous and competitive teacher training programs would be prejudiced in the certification process in comparison with graduates of lesser quality institutions.¹⁰⁵

¹⁰³ Defendants' Proposed Findings of Fact, p. 82.

¹⁰⁴ *E.g.*, As Dr. Thorndike pointed out:

There is abundant evidence that collegiate institutions in the United States vary enormously both in the level of verbal and quantitative ability of the students whom they admit and in the standards of performance that they require in their course work. There is no reason to doubt that this is also true of the colleges of South Carolina. Thorndike written statement submitted at deposition, p. 2, (Ex. P-THO-2).

¹⁰⁵ For example, a graduate with a barely passing average (2.2 grade point average) at an excellent institution might be at least as qualified as a student who obtains a significantly higher average grade (3.5 grade point average) at an inferior school. According to plaintiffs, however, it would be unfair to compare these two graduates against a single criterion assessing their academic performance. To do so would depart from what plaintiffs maintain is the only

In view of these variations in the quality of education of the 25 institutions, the State has a clear interest in ascertaining the *level* of academic preparation in fact achieved by the teacher candidates. The record demonstrates that the fact that a prospective teacher holds an undergraduate degree from a State-approved teacher training institution is insufficient to assure that the candidate possesses the minimal knowledge necessary to teach effectively.¹⁰⁶ The State can reasonably make use of some objective measure by which individual candidates can be fairly evaluated. If properly validated, such a measure assesses the output of the educational process, which in turn reflects both the ability of the prospective teacher and the success of the teacher training program.¹⁰⁷ Plaintiffs confuse the *approval process*, aimed at continuous upgrading of the quality of the teacher training programs, with the *certification process*, aimed at identifying those candidates with the minimal academic knowledge necessary to teach. Both are legitimate governmental objectives; and there can be no reasonable basis for asserting that the State's adoption of an approval process restricts its ability to require achievement of some specified minimal competence by individual candidates. The development of a statewide minimum standard of academic preparation in the ETS validity study, therefore, was an entirely proper element of the study design.

b) *The study design was not required to include direct black/white comparisons among institutions*

Plaintiffs find fault with the study design because it did not provide a means by which to establish (or rebut) their contentions that (1) there is a so-called black teacher education curriculum in South Carolina that is different from a white teacher education curriculum; and (2) the NTE relate less well to the black curriculum

permissible standard, namely, whether the candidate attained a minimal level of the academic knowledge which his teacher training program sought to impart. No State (or court) concerned with the quality of public school education could in good conscience accept such a proposition.

¹⁰⁶ Defendants' Proposed Findings of Fact, p. 81.

¹⁰⁷ Dr. Thorndike elaborated on this point in his deposition. When asked whether a difference in average scores attained by examinees from different institutions would suggest "a difference in the curricula" at the various institutions, Dr. Thorndike replied:

Not necessarily. It could be so interpreted. It might be indicative in the difference in character of the input to the programs in the two types of institutions rather than the curriculum that they received while they were being taught. Thorndike deposition, p. 29 (Ex. P-110).

and are therefore inappropriate tests for blacks.¹⁰⁸ They allege that a study *could have been* designed to serve this purpose and, for that reason, ETS should have done so. Plaintiffs' criticism has two obvious flaws.

First, the fact that there is a difference in the mean NTE scores of candidates attending predominantly black and predominantly white teacher training institutions does not necessarily support a conclusion that the two types of institutions have different curricula. The difference in mean scores might well emanate from the students' different socio-economic backgrounds, different educational opportunities in preparatory schools, and different levels of readiness to do college-level work, as well as the differing resources and programs at the particular institutions.¹⁰⁹ The argument that there is something called "black teacher education" that should be accounted for in the study design simply has no foundation. Plaintiffs may have suggested an interesting research assignment but they have not documented any reason why ETS was required to conduct such an inquiry in the process of performing this validity study for South Carolina.

Second, an indirect measure of the comparison between black and white institutions was available to plaintiffs and investigated by them. Each of the judges was assigned a code number that permitted cross-checking among documents to identify the institution with which the judge was affiliated. Those institutions were classified by the State Department of Education, for purposes of the study, as "predominantly black" or "predominantly white". Analysis of the judgments based upon these racial designations does not support

¹⁰⁸ Plaintiffs' Proposed Findings of Fact, ¶¶ 233-38.

¹⁰⁹ For example, institution X (predominantly black) and institution Y (predominantly white) both prepare prospective chemistry teachers using programs that have been approved by the State. One element of that preparation is the acquisition of knowledge about chemistry, the subject to be taught. The NTE Area Examination in Chemistry, Physics & General Science is administered and the mean score of candidates from institution X is lower than the mean score of candidates from institution Y. Plaintiffs suggest that institution X has a black orientation or approach to teaching chemistry that is not recognized by the test and therefore the test is not appropriate as an alternative measure of the level of achievement of the candidates from institution X. Far more likely, institution X strives to impart exactly the same knowledge of chemistry as institution Y, but it has an open enrollment, its students have an inadequate background in high school chemistry that requires remedial work, and there are insufficient funds for laboratory staff and equipment so students have less opportunity to acquire laboratory-based knowledge.

plaintiffs' contentions. With respect to the content review judgments, judges from predominantly black institutions responded "content appropriate" more frequently than those from predominantly white institutions on six tests¹¹⁰ and less frequently on 18 tests; overall the judges from predominantly black institutions responded content appropriate 67.28% of the time compared with 73.11% for the judges from predominantly white institutions.¹¹¹ Even more significantly, a comparison of the knowledge estimation judgments using these racial designations shows extraordinary similarity between the evaluations of the two categories of judges. The means based on judges from predominantly black institutions were *higher* than the means based on judges from predominantly white institutions on 11 tests and *lower* on 11 tests. Further, the means for all tests for the groups of judges from predominantly black and predominantly white institutions are almost exactly even at 50% (49.23% for black; 50.63% for white).¹¹² Notwithstanding plaintiffs' contentions, therefore, the data show that educators at predominantly black institutions have virtually the same sense as their colleagues from predominantly white schools of what academic knowledge is necessary to graduate from an approved program and to teach effectively in South Carolina. There is no reason to doubt that, within the limitations of available resources and the ability of the students, the programs at the predominantly black schools are designed and operated to meet the standards of the judges from these schools who participated in the study.

c) *The NTE can be used properly for the purpose set out in the validity study*

Plaintiffs contend that the use of the NTE reflected in the validity study is inappropriate because the NTE cannot be used to

¹¹⁰ Interestingly, the judges from black institutions found the area tests in Elementary Education; Chemistry, Physics and General Science; and German to be more content appropriate than did their counterparts from white institutions, indicating there is no pattern as to subject matter in these comparisons.

¹¹¹ Stipulation No. 10 covers the underlying data on which these calculations are based. There are, of course, many factors which might have contributed to this modest difference in overall assessments other than difference in the training program attributable solely to the racial composition of the student body.

¹¹² A table setting out the means of percentage responses made by Knowledge Estimation Panel members from predominantly black institutions and predominantly white institutions for each of the tests is in the record at Manning deposition, written questions on redirect examination, Q 120-22 (Ex. S-8).

check the quality of a particular teacher training program. In many respects, this contention suffers from the same logical and policy objections as the plaintiffs' argument with respect to the development of a statewide standard of minimal competence. It is erroneous, in any event, because it misinterprets evidence in the record regarding the purpose of the NTE and its proper use.

As discussed in Appendix A, the NTE are clearly not designed or developed to serve as a measure of the quality of any individual teacher training program. Quite the contrary is the case; the NTE are designed to provide a *national* norm against which particular programs or their graduates can be evaluated. It is conceded that NTE score results are not being used, or proposed to be used, by South Carolina to establish the content validity of the NTE for a particular college. It does not follow, however, that the NTE cannot be validated for use in South Carolina as an objective measure of the academic preparation achieved by candidates graduating from the different institutions in the State.¹¹³

As plaintiffs emphasize, however, any use of the NTE for evaluating a particular teacher training program should begin with an awareness regarding the extent to which the content of the NTE is related to the college's program. The cautionary comments contained in the 1974 ETS Guidelines cited by plaintiffs are designed only to remind the institution that uses NTE scores for self-evaluation or self-comparison purposes that performance on the NTE will undoubtedly be affected by the extent to which the content of the test matches the particular teacher training program.¹¹⁴ ETS does not advise, however, that the NTE *must* be validated against a particular teacher training program before it can be fairly used for certification purposes with respect to graduates of such programs. Regardless of the content of the teacher training program, the NTE may still provide as to such graduates a fair measure of their academic preparation to be compared with a level determined to be necessary for certification.

Since the validity study provides a basis for concluding that the NTE are a fair measure of the content of the 25 teacher train-

¹¹³ Moreover, it would be entirely appropriate for such institutions to review NTE results to provide some measure of the results of their particular program, and a great many institutions across the nation do use the NTE for this purpose. See discussion in Appendix A at p. A-22.

¹¹⁴ Plaintiffs' Proposed Findings of Fact, ¶ 224.

ing institutions in South Carolina, the judgments of the Knowledge Estimation Panel members can fairly be relied upon to establish a statewide standard of academic preparation as reflected by achieving a designated minimum score on the NTE. Contrary to plaintiffs' assertions, the study provides reliable evidence regarding the validity of the NTE for use in South Carolina.¹¹⁵ When stripped to its essentials, plaintiffs' contention amounts to a generalized attack on the use of any standardized tests for persons with different educational experiences. The argument that each individual can be assessed only against the content of his particular educational program is simply a more sophisticated and verbose version of the basic position of plaintiff National Education Association, namely, to "kill certification tests."¹¹⁶

2. *There Was No Necessity to Relate the Content of the NTE to the Job of Teaching by Means of a Formal Job Analysis*

The plaintiffs maintain that the validity study design was defective because no effort was made by either ETS or South Carolina "to gather data which would directly compare the content of the NTE with the content of the academic knowledge essential for the job of teaching a particular grade level or a particular subject matter area."¹¹⁷ In the absence of such data, which plaintiffs suggest could be obtained only through a formal job analysis, it is charged that the study provides no information "from which one can directly infer that the NTE are a representative sample of the academic knowledge required to be an effective teacher."¹¹⁸ Both aspects of the plaintiffs' argument—the need to relate performance on the NTE directly to the job of teaching and the need for a formal job analysis—find no support in the applicable professional standards or the ample evidence in the record.

To some extent, this contention reiterates plaintiffs' view that the NTE cannot possibly measure the content of the 25 different teacher training programs in South Carolina. This appears to be plaintiffs' only response to the assumption, reflected in the study design and endorsed by expert testimony, that the curriculum offered

¹¹⁵ With respect to the validity of the NTE on a national basis, see Appendix A at pp. A-19 through A-22.

¹¹⁶ Letter from Edelfeldt to Vanderpool, dated Dec. 5, 1975, (Ex. S-EDE-9).

¹¹⁷ Plaintiffs' Proposed Findings of Fact, ¶ 227.

¹¹⁸ *Id.* at ¶ 228.

in these teacher training programs constitutes the knowledge required to be an effective teacher. This assumption, of course, is based both on reliable record evidence and on common sense; South Carolina maintains an extensive administrative bureaucracy to study the educational process within the State and to approve and upgrade 25 separate teaching programs designed and operated to impart knowledge which State officials consider necessary for effective teaching. If the NTE are properly validated against the content of such programs, therefore, no need exists for any further effort to relate the NTE to the knowledge required on the job in particular subject matters or particular grade levels.

In an effort to bolster their contention that the NTE must be compared against the job of teaching, plaintiffs assert that the job of teaching is the "only racially neutral domain which can be considered here."¹¹⁹ Even assuming *arguendo* plaintiffs' assertions regarding the history of segregation in South Carolina educational institutions, the study's validation of the NTE against the content of the teacher training programs does not become either improper or unreliable. The NTE are a professionally developed measurement instrument that provides a means for certifying authorities and others to measure the output of all teacher training programs—whether they are superior programs, inferior programs, predominantly black programs, or predominantly white programs.¹²⁰ Plaintiffs appear to argue, however, that South Carolina has no right to make any comparative judgment based upon the NTE scores of students who have encountered different educational experiences, including exposure to a segregated school system, and may have been accordingly disadvantaged in their capacity to master the curriculum to which they have been exposed.

Contrary to plaintiffs' assertions, the pertinent professional standards and the EEOC Guidelines do not require that the job of teaching be considered the performance domain. This argument reflects plaintiffs' basic position that only a criterion-related validity study was appropriate—a position which ETS believes cannot be supported for the reasons discussed in Section II(B) of this brief. Even assuming that content validity is an appropriate methodology,

¹¹⁹ Plaintiffs Proposed Findings of Fact, ¶ 232.

¹²⁰ Plaintiffs have not challenged the NTE on grounds of racial bias. As documented in Appendix A, the NTE are carefully prepared to avoid such bias and no reliable evidence exists to demonstrate that this objective has not been fully achieved.

plaintiffs maintain that a job analysis is required to determine what knowledge is "essential" and assert that no such job analyses "for the job of teaching any of the subject matter areas or at any of the grade levels involved were performed either prior to or during the conduct of the study—either by ETS or the State."¹²² ETS maintains that no formal job analysis was required for this validity study and that all necessary information regarding the knowledge required to teach effectively was available to the State and fully reflected in its approval of the 25 institutions responsible for training teachers in South Carolina.

Plaintiffs' contention with respect to the lack of a formal job analysis reflects two important concessions. First, plaintiffs apparently concede that the knowledge of what and how one must teach is a necessary prerequisite to effective teaching.¹²³ Second, plaintiffs concede that a job analysis need not examine all facets of the job of teaching but can be limited to determining what *knowledge* is required for effective teaching.¹²⁴ The critical issue, therefore, is

¹²² Plaintiffs' Proposed Findings of Fact, ¶ 244.

¹²³ Plaintiffs' experts had a tendency to be skeptical on this point:

Q. Would you disagree with the opinion that for a teacher, knowledge of the area in which one is going to teach, that is, knowledge of the content one is going to teach, would be relevant to an individual's effectiveness as a teacher?

A. Since you asked me to simply agree or disagree, you don't want me to ask questions to clarify it?

Q. Please ask questions if you need to clarify it.

A. What specific definition do you have in mind in terms of relevance? Because relevance is a term that has meaning to me that means of a statistical validated nature. So, therefore, I would have to say I don't know because I haven't seen the evidence on it.

Q. Well, let's change the word 'relevant' to would be important to an individual's effectiveness as a teacher.

A. I would have to ask the same question about 'important.' Do you mean do I agree that it's been demonstrated to my professional opinion to have been established that it is?

Q. Yes.

A. Again, I would have to say that I don't know because I haven't seen the evidence for it. . . .

Q. Would you disagree with the judgment that knowledge to be gained through academic preparation is useful to teaching?

A. Again, I would have to say I don't know because I think that I don't know of the evidence that exists to validate this particular concept. Bartlett deposition, pp. 95-96 (Ex. P-109).

¹²⁴ Plaintiffs apparently have elected not to rely upon testimony of one of their experts to the effect that a test cannot be justified through content validation unless it is shown to sample all important aspects of a job. Bartlett deposition, pp. 60-61 (Ex. P-109). Even with reference to employment tests, the APA

whether a formal job analysis is necessary to support the study design's assumption that the NTE can be properly content validated by comparing them with the content of the South Carolina teacher training programs.

First, a formal job analysis is not required under all circumstances. A "job analysis" is nothing more than a systematic effort to identify traits, skills or proficiencies that may contribute to success on a particular job. It is particularly useful when a prospective test user is not certain what traits will contribute to success. Thus, the professional guidelines distinctly encourage the use of job analysis in selecting tests to be used in, for example, personnel selection.¹²⁵ Contrary to the testimony of one of plaintiffs' experts, however, the professional guidelines do not make job analysis mandatory if sufficient information for the purposes of the study are available through other means.¹²⁶

STANDARDS themselves specify that the performance domain for which a test is content validated may be "restricted perhaps only to critical, most frequent, or prerequisite work behaviors." APA STANDARDS at 29. The propriety of focusing on, and testing for, a few (or even one) "critical" requirements for job success—even in content validating employment tests—is confirmed in the recently promulgated DIVISION 14 PRINCIPLES. Principle (A)(4) of those dealing with content validity incorporates the statement from the APA STANDARDS dealing with "critical" work behaviors. The preamble to Guideline (A) is even more specific. It provides that a job:

domain need not be inclusive insofar as any larger domain is concerned. By this we mean that *it does not have to cover the entire universe of duties of a particular job or of topics covered in a training course. In fact, there may be many domains in the total content universe for any given job.* DIVISION 14 PRINCIPLES at 9-10 (Emphasis added).

¹²⁵ E.g., APA STANDARDS at 29; Principle (A)(1), DIVISION 14 PRINCIPLES, at 19; *see also* EEOC Guidelines, 29 C.F.R. § 1607.5(a).

¹²⁶ To the extent that the EEOC Guidelines appear to render job analyses mandatory, they do so only with respect to "criterion-related" studies, a type of validation discussed in Section II(B) above. *See* 29 C.F.R. § 1607.5(b)(3). With respect to content validity, the Guidelines provide only that "evidence for content . . . validity should be accompanied by *sufficient information* from job analyses to demonstrate the relevance of the content . . ." 29 C.F.R. § 1607.5(a) (Emphasis added).

Similarly, the APA STANDARDS state that test "users *might define* the performance domain of interest to them . . . in terms of appropriately detailed and comprehensive job analyses." APA STANDARDS at 29 (Emphasis added). Where, however, a test "is represented as having content validity for a job," in contrast with a particular job related skill, the APA STANDARDS do require that "evidence of validity . . . include a complete description of job duties." Standard E 12.4, 46. The DIVISION 14 PRINCIPLES provide only that "[j]ob content domains *should be developed and defined by job analysis.*" Principle (A)(1), DIVISION 14 PRINCIPLES at 10 (Emphasis added).

Second, the study design was predicated on the entirely reasonable assumption that South Carolina's action in approving teacher training programs constituted an informed professional judgment that the programs were designed to impart the knowledge deemed necessary for effective teaching. Contrary to the implication from the testimony of plaintiffs' experts, the development of "job analysis" evidence does not invariably require the application of highly formalized techniques by professional psychologists. The APA STANDARDS, while expressing some preference for a formal job analysis, explicitly recognize that:

The performance domain for published tests might be defined by the pooled judgments of job designers, incumbents, and supervisors. Test users might define the performance domain of interest to them in terms of judgments of similar people in their own organizations. . . . APA STANDARDS at 29 (Emphasis added).

Even more explicitly, the DIVISION 14 PRINCIPLES provide that:

Job content domains should be developed and defined by job analysis, which may be a formal investigation, or the pooled judgments of informed persons such as production engineers, job incumbents, their supervisors, or personnel specialists. DIVISION 14 PRINCIPLES at 10 (Emphasis added).

Thus, while the process of approval of South Carolina's teacher training curricula may not constitute a "formal investigation," there is no reason suggested by plaintiffs' evidence why it does not reflect the application of judgments by professional educators and educational administrators of the type contemplated by both the APA STANDARDS and DIVISION 14 PRINCIPLES.

In this light, plaintiffs' complaint about the lack of a "job analysis" cannot be sustained without attacking the relevance to teaching of the South Carolina curricula themselves. This point was most succinctly made by Dr. Thorndike, an acknowledged dean of the testing and measurement profession:

. . . it would be my judgment that if it is reasonable to accept graduation from any teacher training program as a valid indicator of acceptability to be considered for certification as a teacher, then determination of the extent to which the NTE matches the programs, on the average, is a reasonable basis for evaluating the content validity of the NTE.¹²⁷

¹²⁷ Thorndike written statement submitted at deposition, p. 3 (Ex. P-THO-2).

Dr. Thorndike, it should be noted, did not even remark on the presence or absence of, or even the relevance of, a formal "job analysis."

The anomaly is that plaintiffs have *not* challenged the validity of South Carolina's requirement that prospective teachers possess a degree from an institution with an approved curriculum. They acknowledge the requirement of academic achievement as reflected by completion of the curriculum, but oppose a requirement of academic achievement as measured by a test that has been validated against the content of the curriculum¹²⁸ at least without a formal "job analysis" by a professional psychologist. Such a contention seems both formalistic and untenable.

3. The Use of Teacher Educators Rather than Practicing Teachers on the Knowledge Estimation Panels Was Proper

With specific reference to the Knowledge Estimation Panels, plaintiffs contend that the study design should have used practicing teachers rather than teacher educators as panel members. The point seems to be that only practicing teachers know what knowledge is required to teach effectively.¹²⁹ This criticism was considered and rejected by ETS in the process of reviewing the study design. Reliance on persons who spend their professional careers in teaching future teachers to make the knowledge estimation judgment was entirely appropriate and avoided the distortions which might have resulted from the use of practicing teachers.

The record demonstrates that the panel members were eminently qualified to make the judgments required of them. Many of them did in fact have actual experience of the kind stressed by plaintiffs, either teaching or observing or supervising teachers. Even as to those who did not, however, it is difficult to suggest that educators of teachers are ignorant of what a student will need to know to *teach*, as well as to graduate. While practicing teachers might have

¹²⁸ As pointed out by Dr. Thorndike, the anomaly is heightened by the fact that, given the differences in training programs:

One can make a strong argument that an equal level of competence, as evidenced by a uniform examination, is a more equitable basis for certification than is graduation from any teacher training program, when quite different levels of competence are required to achieve graduation from different colleges. Thorndike written statement submitted at deposition, p. 2 (Ex. P-THO-2).

¹²⁹ Plaintiffs' Proposed Findings of Fact, ¶ 179.

had more direct exposure to what was needed, their prospective on the other hand would almost inevitably be somewhat narrowed by their specialization.¹³⁰ Thus, while the use of practicing teachers was a possible alternative, it was not the only acceptable procedure.¹³¹

Plaintiffs also suggest that faculty members had a systematic tendency to overestimate what they had taught their students and may thus have had a disposition in reviewing the NTE to underestimate the difficulty of some items.¹³² Here again, the suggestion

¹³⁰ When asked whether a successful elementary school teacher would "be able to accurately hypothesize a group of persons with the minimum amount of academic knowledge to teach effectively in the elementary school," Dr. Manning responded:

Mr. Keeling, I have reflected on that particular position quite a bit because it was clearly a consideration that we thought about in conjunction with the design of the study.

* * * *

There are certain problems that I think on further analysis led me to the conviction that the classroom teacher that you have hypothesized may very well not be in the best position to provide judgments of this kind. And I would—so I would cite only two kinds of factors to illustrate this problem.

The first is that the successful elementary school teacher that you are referring to may, indeed, have very much specialized with respect to her experience in a particular school for several years teaching a particular grade. And, that specialization would tend to a degree to remove her or him from appreciation of the minimum qualifications of a beginning teacher in a variety of settings.

And the second is that—I base this not only on the observation of classroom teachers, but also on my own experience as a teacher and as one who has taught numbers of teachers—and that is that there is a very limited opportunity for classroom teachers really to observe other teachers, especially to observe beginning teachers.

And there is, therefore, a tendency for them to be often more removed from the task at hand here—which I continue to reiterate is the task of dealing with beginning, inexperienced teachers, and the certification of the beginning [inexperienced] teacher—to remove them from an opportunity to have the kind of extensive experience with respect to that kind of observation that those who are charged with the responsibility of preparing those beginning teachers do have as a regular course of their responsibilities. Manning deposition, pp. 856-59 (Ex. S-8).

¹³¹ As Dr. Linn observed, "I suggested that that would be an additional procedure that I thought would have added a desirable component to the study, but I don't mean to imply that it's the only procedure." Linn deposition (Vol. II), p. 58 (Ex. P-116). When asked whether other kinds of persons should have been selected as panel members, Dr. Thorndike did not even mention practicing teachers. Thorndike deposition, p. 41 (Ex. P-110).

¹³² Bartlett deposition, p. 69 (Ex. P-109); Thorndike deposition, pp. 40-42 (Ex. P-110).

was in the nature of professional speculation. As Dr. Thorndike candidly conceded:

I couldn't claim to have made observations but I have the impression that teachers typically make their tests too hard for their students, which implies that they think the students have learned more than they have, in fact, learned.¹³³

In any event, ETS' own review of the pertinent literature had suggested the existence of this tendency and ETS took steps to counteract it by developing an appropriate training exercise for the Knowledge Estimation Panel members.¹³⁴ This may well have moderated any tendency to underestimate item difficulty, thus responding to criticisms anticipated by ETS long in advance of plaintiffs' attack on the study.

4. The Use of a Majority Rule Standard for Determining Content Appropriateness Was Reasonable

Plaintiffs contend that the use of a majority-rule standard—i.e., 51% of the panel members making a judgment regarding the content appropriateness of a particular question—was too lenient.¹³⁵ This standard was selected by South Carolina because it embodied the traditional and proper rule that decision on such matters of judgment can be made by a majority of those equipped to express a view. Although other standards might have been selected, plaintiffs have not demonstrated that the choice reflected in the study design was erroneous or improper.

Virtually all the experts who testified on this issue agreed that the decision was to some degree arbitrary.¹³⁶ Recognizing this, plaintiffs' experts in no sense agreed on the level at which the line should be drawn. While Dr. Thorndike suggested that it should be in the

¹³³ Thorndike deposition, p. 40 (Ex. P-110). Dr. Thorndike did indicate that he could think of no other group suitable for making the judgments made by the knowledge estimation panel members without this tendency to overestimate. He considered "the other alternative" to be "supervisor[s] of teachers," *Id.* at 41.

¹³⁴ REPORT, pp. 53-56, 61 (Ex. S-MAN-32B).

¹³⁵ Plaintiffs' Proposed Findings of Fact, ¶ 200.

¹³⁶ *E.g.*, "I think it is a matter of judgment." House deposition, p. 52 (Ex. P-87); "[W]here in the range from 60 to 90 percent, one should set one's standard seems to be quite an arbitrary decision." Thorndike written statement submitted at deposition, p. 5 (Ex. P-THO-2).

range of 60% to 90%, he indicated that he would not in any event attempt to draw the line himself.¹³⁷

Plaintiffs' arguments that a higher standard should have been used overlook two important facts. First, the 51% standard used in the study was, in an operational sense, significantly higher than 51%. For a panel consisting of, for example, eight judges, 51% agreement was, in practice, five out of eight or 62.5%. For a panel of ten judges, 51% agreement required six out of ten or an operational standard of 60%. Second, plaintiffs' contention that one could place "increasing confidence" in the judgment that an item was content appropriate as a higher standard was applied is a two-sided coin. As Dr. Linn pointed out, the other side of the coin is that the use of an increasingly stringent standard would *decrease* the confidence that one could place in a resulting conclusion that the item was "inappropriate." Plaintiffs' preference is simply for a standard that would make the NTE *appear* less appropriate.

In fact, a higher standard, such as 60% (or 66%), would not have altered the overwhelmingly favorable assessments of the Content Review Panels regarding the appropriateness of the NTE for use in South Carolina.¹³⁸ As pointed out earlier, the panel members concluded that at least 90% of the questions on four components of the Common Examinations were content appropriate and that at least 90% of the questions on 13 of the 20 Area Examinations were content appropriate. As reflected in Table 2, increasing the standard to 60% or even 66% has only a minimal effect on these collective judgments.

¹³⁷ "I think if I were doing it, I would provide information on the number of items that met the standard at one level, say 60 percent, and also information on the number of items that met the standard at a somewhat higher level, say 80 percent or 70 percent; and provide to the user both types of information. Having already said it is an arbitrary decision what standard one sets, I am disposed to let somebody else be arbitrary by providing them as much information as possible and not imposing my personal judgments in order to be reasonable in the case." Thorndike deposition, p. 66 (Ex. P-110).

¹³⁸ Plaintiffs' recomputations in support of their argument contain an important error. Their charts treat both "DNK" ("do not know") and "SNK" ("should not know") responses as the equivalent of judgments of content "inappropriate." The DNK responses are not judgments of content inappropriateness, as is discussed *infra* at pp. 105-08. The SNK designation was used by ETS to account for clerical errors in the test documents (REPORT pp. 64-66, Ex. S-MAN-32B) and cannot be the equivalent of a judgment of content inappropriateness.

TABLE 2

Percentages of Questions Judged Content Appropriate, Not Counting
Do Not Know Responses, At Various Levels of Agreement Among Judges

Common Examinations	Percent of Items Appropriate at 51% (per study, Table 17, p. 73)	Percent of Items Appropriate at 60%	Percent of Items Appropriate at 66%
Professional Education	83.1%	79.0%	74.4%
Written English Expression	96.6%	95.6%	92.2%
Social Studies Literature & Fine Arts	98.3%	98.3%	98.3%
Science	63.2%	58.6%	46.9%
Mathematics	91.6%	86.7%	85.0%
	92.5%	92.5%	92.5%
<u>Area Tests</u>			
Agriculture	96.0%	96.0%	96.0%
Art Education	73.0%	65.4%	65.0%
Art Education	73.0%	65.4%	65.0%
Biology & General Science	95.0%	91.2%	87.7%
Business Education	96.2%	94.6%	90.0%
Chemistry, Physics & General Science	94.0%	90.0%	81.6%
Early Childhood education	96.4%	96.4%	96.4%
Education in Elementary School	91.2%	87.6%	80.9%
Education of the Mentally retarded	96.7%	96.7%	95.3%
English language & Literature	78.6%	75.2%	69.9%
French	95.9%	93.1%	88.0%
German	61.8%	61.8%	61.8%
Home Economics Education	97.6%	97.6%	96.0%
Industrial Arts Education	77.8%	74.4%	69.2%
Mathematics	95.4%	92.5%	90.0%
Media Specialist	95.2%	95.2%	93.0%
Music Education	87.0%	81.9%	80.6%
Physical Education	94.0%	91.6%	85.0%
Social Studies	95.6%	93.2%	88.6%
Spanish	88.4%	84.4%	76.4%
Speech-Communication & Theatre	72.6%	72.6%	62.5%

There is virtually no change in the classifications as to content appropriateness of the six subtests of the Commons when the 51% requirement is increased to either 60% or 66% and the 90% requirement is reduced to 85%:

- Number of subtests for which 90% or more of the items were judged content appropriate by 51% of the responding judges: *four*
- Number of subtests for which 90% or more of the items were judged content appropriate by 60% of the responding judges: *three*
- Number of subtests for which 90% or more of the items were judged content appropriate by 66% of the responding judges: *three*
- Number of subtests for which 85% or more of the items were judged content appropriate by 51% of the responsible judges: *four*
- Number of subtests for which 85% or more of the items were judged content appropriate by 60% of the responding judges: *four*
- Number of subtests for which 85% or more of the items were judged content appropriate by 66% of the judges: *four*

More variation in the initial classification is produced with respect to the Area Examinations studied under the various proposed standards, although even here the results are also quite close:

- Number of area tests for which 90% or more of the items were judged content appropriate by 51% or more of the responding judges: *thirteen*
- Number of area tests for which 90% or more of the items were judged content appropriate by 60% or more of the responding judges: *twelve*
- Number of area tests for which 90% or more of the items were judged content appropriate by 66% or more of the responding judges: *seven*
- Number of area tests for which 85% or more of the items were judged content appropriate by 51% or more of the responding judges: *fifteen*
- Number of area tests for which 85% or more of the items were judged content appropriate by 60% of the responding judges: *thirteen*

- Number of area tests for which 85% or more of the items were judged content appropriate by 66% or more of the responding judges: *eleven*

But plaintiffs' reliance on a variation in the number of tests found to be *per se* content appropriate under various standards is misplaced. The *per se* rule is not the end of the analysis and a variation in the initial classification does not necessarily affect the outcome of the study.

There were large quantities of data produced by the study and these data had to be organized and analyzed to permit a classification of the tests into those that were appropriate for use in South Carolina and those that were not. As the researchers who conducted the study recognized:

There is no single, professionally established means of interpreting the data resulting from this study. Professional judgments must be made about the meaning of the data in terms of the teacher education programs in South Carolina and about the standards to be established for use of the test scores.¹³⁹

The ETS research team approached the problem of applying their best professional judgment to the data by establishing a *per se* rule that took account of the four separate judgments about content congruence made by the judges. If the data with respect to a test exceeded the standard on all four measures the test was *per se* content appropriate. The standard is summarized on Table 3. If the test was not *per se* content appropriate, the data were subjected to "special review" by the ETS researchers and a judgment was made about what the data meant. After that "special review," the tests were classified as "very closely related," "closely related," "reasonably related" or "least related" to the South Carolina curriculum. A number of tests that did not meet the *per se* standard were classified as "very closely related" or "closely related" on the basis of the ETS judgment as to the proper interpretation of the data. The basis for the ETS judgments with respect to each of the tests is set out in the REPORT at pp. 151-66. Plaintiffs do not challenge any of those judgments as improper or not supported by the data. Their challenge of the level at which the *per se* rule operates does not affect the rationality of these judgments or the outcome of the study.

¹³⁹ REPORT, p. 141 (Ex. S-MAN-32B).

TABLE 3 *

Criteria for Performing Detailed Review of Correspondence
between Test Content and Program Content

Aspect of Correspondence	Measure of Correspondence	Theoretical Range	Observed Range	Standard for Requiring Special Review
Content appropriateness of questions	Percentage of questions classified as content appropriate by total panel (mean for two test editions)	0% to 100%	62% to 98%	Less than 90%
Coverage of test	Number of content areas whose non-inclusion in the test was noted by two or more judges	0 to total number of content areas in test	0 to 2	1 or more
Relative emphasis in test and curriculum	Index of difference in relative emphasis	0 to 100	9 to 61	More than 40
Overall similarity of test and teacher education sequence	Ratio of number of judgments that test closely parallels or does not differ appreciably from curriculum to total number of judgments about similarity	0% to 100%	0/3 to 100%	Less than 1/2

*A similar table appears at page 142 of the REPORT.

5. *The Lack of a Pilot Study Provides No Basis for Discarding the Results of the Validity Study*

Plaintiffs complain that ETS should have conducted a pilot study before finalizing the study design. According to plaintiffs, such a pilot study would have revealed the procedural deficiencies in the study's design which plaintiffs contend resulted in unreliable data.¹⁴⁰ This criticism is not supported by the record.

Although it might have been desirable to conduct a pilot study, the failure to do so does not provide any independent basis for challenging the validity study. This criticism is symptomatic of the tendency of professionals in any area to believe that they could have improved upon another's effort—given the benefit of hindsight and free of any constraints of time and resources. The issue is not whether a pilot study should have been conducted or whether the study might have been improved in one of the many respects suggested by the experts; the issue rather is whether any of the substantive criticisms of the study design presented by plaintiffs provide a basis for disregarding the study's results.

The ETS decision to forego a pilot study was entirely appropriate under the circumstances. South Carolina was indeed under considerable pressure, originating from both ETS and the plaintiffs to reexamine its certification procedures and bring them into compliance with ETS Guidelines and legal requirements. Plaintiffs maintain that the pressures operating upon both ETS and the State "provided an incentive for them to bias the study toward a result that the NTE are 'content appropriate' and that a minimum composite cutoff score equal to or higher than 975 is justified."¹⁴¹ One can only imagine the contention that would have been advanced by plaintiffs if South Carolina had failed to move promptly in 1975 to examine the legitimacy of its continued use of the NTE. It is certainly true that this validity study was not conducted in a vacuum; it was performed by ETS on contract for South Carolina under exceptional pressures resulting from limited time and resources and the anticipated vigorous challenge in this litigation to any validity study relied upon by the State. Plaintiffs offer no persuasive reason why these practical considerations should be ignored by the court in evaluating the reasonableness of the ETS decision not to conduct a pilot study.

¹⁴⁰ Plaintiffs' Proposed Findings of Fact, ¶ 152.

¹⁴¹ *Id.* at ¶ 208.

**B. Plaintiffs' Criticisms of the Execution of the Study Design
Involve Alternatives as to which Reasonable Choices Were
Made**

Plaintiffs' criticisms of the study's execution are obviously advanced in the hope that their cumulative effect would be to disqualify the study. Zero plus zero, however, still equals zero. In fact, each of the criticisms offered by plaintiffs can be rebutted on one of three grounds: (1) it is an inaccurate conclusion from the data; (2) the effect on the score levels set by the study is either insignificant or such that a more *lenient* certification standard is created; or (3) expert opinion with respect to the practice or method differs and the decision of the ETS professional team was reasonable. This section discusses these criticisms of the execution of the study design in four subsections: an assessment of the quality of plaintiffs' expert testimony and their depositions of participating panel members; discussion of criticisms relevant to both knowledge estimation and content review panels; and then discussion of the issues relevant to each of the two kinds of panels.

**1. The General Quality of Plaintiffs' Evidence Is Such that
It Is Entitled to Little Weight**

Plaintiffs offered the opinions of experts and the testimony of 81 judges who participated in the study. There are important limitations that affect the quality of the evidence from both sources.

a) Testimony of experts

Plaintiffs did considerable searching to find experts with views that were helpful to their case. After receiving the validity study on January 7, 1976, and prior to answering defendants' interrogatories on April 5, 1976, plaintiffs contacted at least nine experts with reference to the study.¹⁴² Of this group, only four, Dr. Medley, Dr. House, Dr. Bartlett, and Dr. Goldstein, were called to testify.¹⁴³ These witnesses found their views to be in accord with plaintiffs on virtually every issue relevant to this case. This is not surprising since the measurement field, like most other professional fields, has ardent advocates of conflicting points of view. Dr. House, for example, is a well-known opponent of standardized tests who has been

¹⁴² Plaintiff-Intervenors' Response to Defendants' Second Interrogatories, Int. No. 105, at 124-25. In addition to those experts called as witnesses, plaintiffs evidently contacted a Dr. Ralph Tyler, a Dr. Marvin Dunnette, a Prof. Kovacs (possibly with NEA) and a Mr. Charles Offutt.

¹⁴³ Dr. Thorndike, an eminent expert in testing and measurement, was called by the plaintiffs and defendants.

employed previously in an expert capacity by plaintiff National Educational Association.¹⁴⁴ Dr. Bartlett has written extensively about, and is still a proponent of, differential validity despite the substantial professional view to the contrary. *See, e.g.*, p. 35 fn. 61, *supra*. Dr. Medley, an ex-ETS employee, had taken a position in 1972 that one portion of the Common Examinations was disadvantageous to blacks. *See* discussion in Appendix A, pp. A-29 through A-32. Interestingly, each of plaintiffs' experts testified only to his own view. None was asked whether his view represented a majority or minority opinion within the profession.

The opinions of these experts should also be treated with some skepticism because of the quality of their advice. Each rejected the obvious practical limitations of present technology, time and money when espousing alternatives to the ETS design. South Carolina, of course, cannot afford to wait until the state-of-the-art is sufficiently advanced to permit the most graceful design or the most perfect results; nor can the State afford to commit unlimited resources to the resolution of this research problem. Undeterred by such mundane considerations, plaintiffs' experts liberally opined regarding what might be accomplished at some unspecified time in the future with unspecified amounts of time and manpower. They were not asked what could realistically be accomplished now with limited resources—in short, to view the task as ETS was required to do in performing this validity study for the State.

b) *Testimony of judges*

Plaintiffs amassed over 2,700 pages of transcript from the depositions of 81 of the 456 faculty members who participated in the ETS validity study as judges.¹⁴⁵ They have offered in evidence what purport to be summaries of the depositions and the expert testimony of Dr. Irwin Goldstein with respect to these summaries.¹⁴⁶

¹⁴⁴ As discussed more fully below, Dr. House has previously been hired by NEA and a local NEA affiliate to review an educational program, involving standardized tests, of which NEA was critical. *See* footnotes 269, 270 and 271, *infra*; House deposition, p. 84 (Ex. P-87). There, as here, Dr. House's professional opinions were consistent with the outcome sought by NEA.

¹⁴⁵ Plaintiffs describe the selection process for these deponents in their Proposed Findings of Fact, ¶¶ 153-158.

¹⁴⁶ Plaintiffs summaries are Exhibit P-134. The parties have stipulated into evidence the mechanical steps performed by the United States in identifying the sample of judges to be deposed. Stipulation No. 9 (Ex. P-128). However, there is no testimony that the steps set forth in that stipulation constitute a "random" sampling. The record does contain Dr. Goldstein's opinion that the procedures followed by plaintiffs would have led to a random selection. But Dr.

Plaintiffs attempt thereby to disqualify the validity study by showing that some proportion of the judges (1) did not prepare for the tasks they were asked to undertake; (2) did not know which category of examinations (Area or Commons) they were reviewing; and (3) did not understand or follow the instructions.

The data and examples extracted by plaintiffs from those depositions (and contained in their deposition summaries) are reliable only if the circumstances under which the data were developed were fairly designed to elicit accurate information. Plaintiffs' own expert expressed some basic reservations on this score. Indeed, Dr. Goldstein pointed out that eliciting this information through depositions rather than informal interviews was undesirable,¹⁴⁷ that the accuracy of the data could have been affected if the judges were under stress, and that obtaining the data through depositions could have placed the judges under stress.¹⁴⁸ The deficiencies inherent in these depositions recognized by Dr. Goldstein were compounded by the way in which they were conducted in at least three important respects: the timing of the depositions; the extensive use of leading questions; and the failure to permit the witnesses to inspect any of the relevant documents during their testimony about those documents.

Goldstein conceded that he had "no personal knowledge of the process" by which they were selected and that his opinion was based solely on the process as "described to me." Goldstein deposition, p. 74 (Ex. P-140). Indeed, Dr. Goldstein was asked, for purposes of soliciting his opinion as to the randomness of the sample, to "assume that the panel members were in fact selected" in the way that had been described to him (*id.* at 64, emphasis added), and he went out of his way to stress on cross-examination that he had no first-hand knowledge of the process. *Id.* at 74-75. Stipulation No. 9 itself was not shown to Dr. Goldstein. Consequently, there is no basis whatsoever in the record for concluding that the process "as described" to Dr. Goldstein was the same as the process to which the parties stipulated. There is therefore no way of knowing whether the sample was in fact "random."

It is useful to note that this "random" sample happened to produce the heaviest concentration of judges in the groups that examined Test C of the Commons (social studies, literature and fine arts.) The "random sample" included six of the nine judges who served on the content review panel for the social studies portion and six of the 12 content review judges for the literature and fine arts portion, although no other content review panel was represented in the sample by anywhere near the many judges. Test C was the single test most criticized by plaintiffs' experts as inappropriate for use in South Carolina. See Appendix A, pp. A-23 through A-38.

¹⁴⁷ Goldstein deposition, p. 73 (Ex. P-140).

¹⁴⁸ *Id.* at 79-80.

(i) *timing*: The judges' participation in the study was completed during the period October 31 to November 5, 1976.¹⁴⁹ Plaintiffs were served with a copy of the study on January 7, 1976, one day after they were ordered to complete their discovery with dispatch. Plaintiffs were able to mobilize their discovery effort with respect to the 81 judges within one week,¹⁵⁰ but delayed until late in the discovery period permitted by the court before undertaking the depositions.¹⁵¹ Once the notices of depositions were served on the State, the plaintiffs moved to complete them with great haste; often three or more depositions in each of two locations were scheduled for a single day;¹⁵² and all of the depositions were conducted using a standard form outline of questions. The 81 depositions were substantially completed in three weeks.¹⁵³

Plaintiffs' strategy of waiting until the very end of the discovery period and then proceeding through 81 depositions on a crash schedule had three calculated effects. First, the witnesses' memory of the detailed task they had performed, involving as many as 320 different test questions, on one day in late October or early November had faded to the maximum extent possible.¹⁵⁴ Second, the

¹⁴⁹ REPORT, p. 60 (Ex. S-MAN-32B).

¹⁵⁰ The names and school affiliations of most of the judges were available at the time the REPORT was released by the State. They are listed at pp. D-1 through D-40, Appendix D (Ex. S-MAN-32A).

¹⁵¹ The first deposition, of Walter J. Fraser, Jr., was taken on March 8, 1976, in Charleston. At this time, discovery was scheduled to end on March 29, 1976 pursuant to Order No. 4 of this court. After the depositions of the judges were noticed and mostly completed, this schedule was extended with respect to particular parts of the discovery. Order No. 5.

¹⁵² For example, on March 18, 1976, plaintiffs took the depositions of Carol Von Kleist and Keith Wagner in Newberry; Lillie Glover in Orangeburg; and Annabelle Boykin, David Franklin and Kathryn Powell in Rock Hill.

¹⁵³ Only three depositions were taken in April: Kenneth Creel (April 5, 1976) (Ex. P-43), John Whitaker (April 5, 1976) (P-38), and John Tribble (April 16, 1976) (Ex. P-40).

¹⁵⁴ Dr. Goldstein acknowledged the lapse of time would affect the deponent's recollection and that this in turn would affect the confidence one could place in the witness' statements. Goldstein deposition, pp. 78, 101 (Ex. P-140). A number of the judges commented on this problem. For example:

Q: Would you have had the same impression at the time of your participation on the panel?

A: No, I would have had more knowledge about it then.

Q: But at this point you don't remember exactly?

A: No, that's right.

Braun deposition, p. 16 (Ex. P-12).

State did not have the time or the resources to prepare the witnesses even to the extent of having them review the study documents reflecting their work or the REPORT. Third, because the depositions were often scheduled simultaneously in different locations the State was unable to have its trial counsel, who were familiar with the issues, attend all of the depositions and lawyers with no experience on the case had to be pressed into service.

Plaintiffs' strategy was apparently successful to some degree. Judges who had attended the one-day sessions in the fall, studied written instructions, listened to oral instructions and familiarized themselves with several documents, had forgotten some of the details when they were deposed four months later. Plaintiffs' contentions that the judges did not understand or follow the instructions must be evaluated in light of plaintiffs' tactics and the familiar imperfections in human memory reflected in these depositions.

(ii) *leading questions*: Plaintiffs' counsel made extensive use of leading questions going to what they had identified as the primary issues in the case.¹⁵⁵ For example, on the issue of the comparisons made in the content review task:

In making determinations in relation to paragraphs 1, 2, and 3 on page E-6, did you first determine the percentage emphasis a content topic received in the curriculum at South Carolina State College and then compare that percentage with the topic's emphasis on the test?¹⁵⁶

and on the issue of the accuracy of the knowledge estimation judgments:

Keeping in mind the instructions you had been given, had been provided you, on pages E-38 and 39, is the curriculum at South Carolina State College so well defined that you could have determined whether ninety percent of the students would have an opportunity to learn the answers to a particular question?¹⁵⁷

and also

In making your judgment on a particular question, as to whether 90% of the students would have had an opportunity

¹⁵⁵ Many of the leading questions are identified in the objections filed by the State with respect to these 81 depositions. The plaintiffs' use of leading questions was in direct conflict with Rule 611(c) of the Federal Rules of Evidence.

¹⁵⁶ Morgan deposition, p. 19 (Ex. P-4).

¹⁵⁷ Jones deposition, p. 20 (Ex. P-45).

to, uh, the, learn the answer to that particular question, uh, did you consider, or did you take into account the fact that, uh, uh, at, student seeking a sufficient amount of Spanish for a certificate, uh, might not have taken that course, where the subject matter, covered by the question was taught.¹⁵⁸

Counsel for the State did not object to the leading form of these questions at the time they were posed and under the stipulation generally entered at the outset of this series of depositions, those objections were probably waived.¹⁵⁹ However, the manner in which plaintiffs' counsel interrogated these witnesses can be considered in determining what *weight* to give to this testimony even if objections as to its admissibility are foreclosed.

Counsel for plaintiffs knew that each of these witnesses was inexperienced; most had never testified or been subpoenaed before in any kind of a proceeding. It was also readily apparent that the State had not interviewed these witnesses in advance of their depositions regarding the nature of their participation in the study. Under these circumstances, the choice of plaintiffs' counsel to rely heavily on leading questions directed at the primary issues in the case was particularly critical to the quality of the data now offered in evidence against the validity study. This strategy permitted plaintiffs to insure, to the maximum extent possible, that the judges' responses would be favorable to their theory of the case, and to fit the judges' responses into the format of the summaries they planned to prepare and submit into evidence.

(iii) *withholding of documents*: Each of the panel members deposed by the plaintiffs was asked about his or her judgments made as part of the study. Those judgments were reflected in several types of documents.¹⁶⁰ All of these documents were available

¹⁵⁸ Platt deposition, p. 30 (Ex. P-24).

¹⁵⁹ The stipulation entered at the outset of the depositions of the judges generally incorporated by reference the stipulation entered previously at the deposition of John Maynard. (*E.g.*, Jeremias deposition p. 1, Ex. P-41). That stipulation provided that all objections would be reserved except objections as to the form of the questions. Maynard deposition, pp. 2-4 (Ex. P-91).

¹⁶⁰ Each of the judges who participated on a Content Review Panel completed two forms, a two-page "Test Content Description Review Form" and a "Question Review Form" of up to five pages, and each knowledge estimation panelist completed a single form that was as much as seven pages in length. Examples of these documents are set out in the REPORT, pp. E-1 through E-7, E-40 through E-44, G-7 through G-13 (Ex. S-MAN-32A). Each form completed by a judge had the judge's name on it and was marked with a code number identifying the judge.

to plaintiffs at the time the 81 depositions were taken. They had been produced by ETS for inspection by the plaintiffs in January 1976 and were marked as exhibits on February 26, 1976 after the appropriate stipulation was entered protecting confidentiality.¹⁶¹ Plaintiffs' counsel never produced these documents for the inspection of the witnesses. If the judges had been able to review the test questions which they evaluated and their specific recorded judgments, their responses to plaintiffs' examination would undoubtedly have been more accurate and precise, even if less favorable to plaintiffs' theory of the case.

More important, plaintiffs' counsel never advised any of these witnesses of the relevant provision of Rule 26(b)(3), Federal Rules of Civil Procedure:

Upon request, a person not a party may obtain without the required showing a statement concerning the action or its subject matter previously made by that person. If the request is refused, the person may move for a court order. The provisions of Rule 37(a)(4) apply to the award of expenses incurred in relation to the motion.

It was readily apparent to plaintiffs that none of these witnesses had the relevant documents in his or her possession.¹⁶² None was represented by counsel and none was likely to know that he or she was entitled to look at the documents reflecting the judgments made or the test questions about which those judgments were made *before* answering *any* of plaintiffs' questions.¹⁶³ It was also readily apparent that the judges would have preferred to have been able to review those documents prior to testifying, had they known of their right to do so. For example,

¹⁶¹ Confidentiality was necessary to honor ETS' representations to the panel members that their individual judgments would not be made public and to ensure that similar research projects in the future would be able to secure the kind of cooperation from educators that was evident in the ETS study.

¹⁶² The forms that the judges completed were taken to ETS for data processing. The documents containing the test questions were carefully controlled by ETS (and retrieved from the judges when they finished their work) for security reasons.

¹⁶³ Plaintiffs were anxious that no person other than counsel for the parties attend these depositions. When counsel for ETS attempted to observe one of the depositions, the Department of Justice took (and maintained for 18 hours) the position that the deposition was *closed* to anyone not personally invited by the Department and refused to continue questioning unless counsel for ETS left the room. Cannon deposition, pp. 15-20 (Ex. P-6).

A: . . . What bothers me is not having my data [the completed forms] in front of me when, you know, on which to respond. I suppose, I suppose it was somewhat subjective when, subjective in that you take a series of percentages, you compare them with another group of percentages, and then you decide what's appreciable and what's not appreciable, and not having my percentages in front of me, I, you know, I could not get the hard and fast answer . . .¹⁶⁴

and similarly,

A: . . . I could only make the answer more specific if you had a copy of my examination, with you, film [visual arts] in particular.¹⁶⁵

By withholding these documents, plaintiffs may have accumulated more data favorable to their theories about the validity study, but only at the cost of diminished quality and reliability.

2. The Criticisms Applicable to Both Panels Are Based Only on Speculation

Plaintiffs have four general criticisms of the execution of the study that apply to the operation of both the content review and the knowledge estimation panels: the judges' teaching experience in the subjects about which they were making judgments; the advance preparation by the judges; the use of pre-specified response forms; and the provision of correct answers on the judges' forms. Each of these points was considered by ETS at the time the study was carried out and none are substantial.

a) Teaching experience of judges at college level

Plaintiffs question the ability of panel members to evaluate questions in areas which they have not taught in teacher training programs.¹⁶⁶ If this amounts to a contention that no question on the NTE should be assessed for any purpose by someone other than a person who had taught the particular subject, it is patently im-

¹⁶⁴ Price deposition, p. 33 (Ex. P-29).

¹⁶⁵ Calhoun deposition, p. 46 (Ex. P-1).

¹⁶⁶ Plaintiffs' Proposed Findings of Fact, ¶¶ 181-82, 192-95. They emphasize calculations which show that the KEP members taught only 39.6% of the items they were asked to review, *Id.* at ¶ 181, and argue from these data that the CRP members had also never taught most of the items evaluated by them. *Id.* at 194.

practical and ignores the procedures built into the study by ETS to handle this problem.

ETS concedes that not every judge had teaching experience with respect to every subject covered by the test questions about which he or she was asked to make judgments. In social studies, for example, which consists of history, economics, government, geography, sociology and anthropology, some judges who taught in social studies programs or departments had taught one or more but not all of these subjects. Similarly, in the science area where one of the tests covers biology and general science (elementary chemistry, physics, astronomy, geology, and meteorology), and another test covers chemistry, physics and general science, some faculty members taught courses in one area but not others. To have required every question to be judged by a person who had actually taught the material would have inflated the number of judges needed to an extent that the study probably could not have been conducted, and certainly would not have been completed within the constraints of time and money under which the State was operating.

ETS took account of this problem in two ways: (1) the preliminary instructions asked the judges to prepare, if they needed to, by inquiring of colleagues, reviewing catalogs or departmental offerings, and consulting other sources; and (2) the judgment forms provided an option for the judges to indicate "do not know" as their response.

Many of the judges came from small institutions or small departments within larger institutions. It was to be expected that such faculty members would be quite familiar with the subject matter taught by their colleagues. Planning within such institutions or departments, with respect to course requirements, catalog descriptions, sequences for majors, and research generally involve extensive discussions and review of written course descriptions that give faculty members a good overview of what is being taught. Further, in planning their own courses, faculty members often have to take into account what is taught in other courses so that the program will mesh to the required extent.

Under these circumstances ETS reasonably believed its solution to be acceptable and practical and plaintiffs have not demonstrated otherwise.

b) *Advance preparation by judges*

The ETS instructions sent to all judges before the panel meetings contained both a description of their assignment and the following suggestions with respect to possible preparation for that assignment:

Before attending the panel session, you may want to draw upon local sources of information regarding the curriculum at your institution, such as your college catalog, specialists in curriculum planning, or other sources available to you. You may also find it helpful to talk with colleagues who have taught specific courses that you have not been called upon to teach.¹⁶⁷

Plaintiffs contend that the study is unreliable because many of the panel members did not in fact undertake any of the suggested preparation.¹⁶⁸

Plaintiffs fail to recognize that this suggestion was made by ETS in the interest of providing the most complete possible participation by the judges. Judges who felt in need of preparation would know that the study design permitted it, and would be encouraged to look to the relevant sources. The most logical inference from plaintiffs' data is that some judges did not take specific steps to prepare because their extensive teaching and faculty experience provided preparation enough. Indeed, one judge said pointedly:

Q: And did you spend any time in preparing for the panel assembly prior to going to Columbia?

A: No. Well, in a sense, yes. My whole life, my whole career, graduate school. Yes. I've spent years preparing. It was all in preparation for this.¹⁶⁹

c) *Use of pre-specified response forms*

All but one of the tasks performed by the Content Review Panel members were presented to them in a manner that required a re-

¹⁶⁷ REPORT, pp. F-34 and E-35 (content review panels), G-1 and G-2 (knowledge estimation panels) (Ex. S-MAN-22A).

¹⁶⁸ Plaintiffs' Proposed Findings of Fact, ¶¶ 180, 188.

¹⁶⁹ Spove deposition, p. 7 (Ex. P-42). Dr. Spove's views on this question were not reflected in plaintiffs' summaries. Indeed, plaintiffs did not ask the judges why they had not specifically prepared. They pursued their questioning only so far as was necessary to establish on the record, if they could, that a particular judge had not done any specific preparation.

sponse in what is known as a "pre-specified" format. For example, in judging the topic by topic similarity or dissimilarity between the NTE and the curriculum of their institution, the judges were given forms that describe the topics covered by the NTE test they were reviewing, and the percentage of that test that was devoted to each topic.¹⁷⁰ The judges were then asked to specify whether the emphasis given to that topic in their curriculum was "the same as," "more than," or "less than" the emphasis given on the test they were reviewing.¹⁷¹ Some of plaintiffs' experts felt that a "free response" rather than "pre-specified" format should have been used¹⁷² and plaintiffs have hesitantly endorsed their experts' views on this technical issue.¹⁷³

Either a "free response" or a "pre-specified response" form is professionally acceptable and in fact both are used in the study. The form on which the content review panel members were asked to list the topics omitted from the tests is a "free response" form. ETS chose a pre-specified response form for other aspects of the study because it facilitated compilation and comparison of the results.¹⁷⁴

d) *Provision of correct answers*

Plaintiffs also challenge the procedure whereby the test questions provided to the panel members had the correct answer marked.¹⁷⁵ The alternative apparently preferred by plaintiffs would give the judges only the questions and the four or five alternative

¹⁷⁰ The "Test Content Description" forms, which describe the topics and percentage emphasis for each test, are set forth in the REPORT at pp. E-7-33 (Ex. S-MAN-32A).

¹⁷¹ Specimens of the form on which the judges were to record their determination on this task are also included in the REPORT at pp. E-2, E-5 (Ex. S-MAN-32A).

¹⁷² Under a "free response" format, the judges would either have been (a) asked to specify the topics covered in their curriculum and the percentage devoted to each, or (b) given the topics covered on the NTE but not the percentage devoted to each topic, and asked to specify the percentage of their curriculum devoted to each topic.

¹⁷³ Although conceding that the experts "are in dispute" on this issue, plaintiffs include it in their catchall list of criticisms of the study and suggest that the use of the pre-specified format may have encouraged the content review panel members to conclude that their respective curricula had the "same" emphasis as the test being judged. Plaintiffs' Proposed Findings of Fact, ¶ 205(a).

¹⁷⁴ REPORT, p. 47 (Ex. S-MAN-32B).

¹⁷⁵ Plaintiffs' Proposed Findings of Fact, ¶ 205(c).

answers so that they would have to select the correct answer for themselves. After considering this possibility, ETS decided that the correct answers should be provided in order to obtain the most accurate results, because this procedure ensured that each judge made his judgment based on comparable information.

A decision not to provide the correct answers usually reflects a concern that if the judges have the correct answers they will mistakenly think they know the answer and consequently will judge the question to be easier than it really is. That theory has an equally persuasive counterpart: if the judges do have the correct answers in front of them it is just as likely that they will recognize answers that they did not know and judge the questions accordingly. Whereas, if the judges do not have the correct answer in front of them, some might pick a wrong answer and think the question is easier than they would if they recognized that they did not know the right answer.

ETS decided that the interest in having each judge use the same information, and the likelihood that there was no adverse affect from providing the correct answer, was sufficient basis for providing the correct answer. Plaintiffs have not provided any research data that indicate otherwise.

3. The Criticisms Applicable Only to the Content Review Panels Are Very Limited in Scope and Effect

Plaintiffs found relatively little to criticize with respect to the Content Review Panels. They contend that the test content descriptions were too broad and allege that some of the judges did not follow their instructions. Neither of those points has support in the record and the criticisms are minor indeed when compared to the very convincing findings of content appropriateness emerging from these panels.

a) Quality of test content descriptions

ETS used test content descriptions for one part of the content review process. These documents contained a listing of the major topics covered by the test and percentage figures indicating the relative emphasis on each topic in the test. They were, in all but a few cases, the actual descriptions used as part of the specifications for the tests.¹⁷⁶ The judges were asked to compare the percentage empha-

¹⁷⁶ REPORT, p. 46 (Ex. S-MAN-32B).

sis in their curriculum with the percentage emphasis on the test and then to indicate whether any major topics were omitted on the test that were included in their curriculum. This latter information provided a measure of the coverage of the test.

Plaintiffs' complaint is that the test content descriptions, which are reproduced in the REPORT, were so broad as to make it unlikely that the Content Review Panel members could identify subjects not covered by the NTE.¹⁷⁷ It would be expected that a test developed against a national norm would not cover *all* topics covered in the curriculum of a particular state. And, if the content description forms had in fact been so broad as to lead the judges to conclude that nothing had been omitted, there might be some substance to this claim. However, the evidence is to the contrary. One-fifth of the Content Review Panel members, 49 in all, listed at least one topic omitted from the test they reviewed. At least one topic was listed as omitted from five of the six sections of the Common Examinations, and from 17 of the 20 Area Examinations.¹⁷⁸

The test content descriptions are realistic, workable documents that were designed to strike a middle ground with respect to specificity. They could have been made more specific. However, the most reasonable inference to be drawn from the fact that the judges found so few topics missing is that ETS has done a thorough job of surveying teacher training curricula and has designed a test that fairly reflects those curricula.¹⁷⁹

b) *Ability of judges to follow instructions*

Plaintiffs' summaries purport to show that 10 of the 41 Content Review Panel members failed to follow the instructions.¹⁸⁰ The summaries distort the evidence and plaintiffs' conclusions are unsupported.

¹⁷⁷ Plaintiffs' Proposed Findings of Fact, ¶ 205(b).

¹⁷⁸ REPORT, Table 18, pp. 75-77 (Ex. S-MAN-32B). Indeed, at least one topic was listed as omitted from the very test for which the test content description, in Dr. House's view, exemplified the undue generality of the descriptions. Compare House deposition, p. 50 (Ex. P-87) and page E-10 of the Appendix to the REPORT (Ex. S-MAN-32A) with p. 75 of the REPORT (Ex. S-MAN-32B).

¹⁷⁹ See Appendix A for a description of these surveys and the test development process.

¹⁸⁰ Plaintiffs' Proposed Findings of Fact, ¶¶ 190-91.

For example, plaintiffs claim that four Common Examinations content review panelists had in mind students majoring in their area of expertise, rather than all teacher education majors. However, an analysis of the transcript indicates otherwise:

Reitzer: The testimony of Dr. Reitzer, who reviewed the social studies portion of the Commons, indicates that he had both teacher education majors and social studies majors in mind. He testified as follows:

Q: Did you have in mind students enrolled in particular fields or areas of concentration?

A: That is correct.

Q: Which fields or areas of concentration?

A: Judging on the percentage of majors in the various fields who indicate *teaching as an area of concentration*. And those who are to be certified in Social Studies beyond just the major area.¹⁸¹

Hall: Mrs. Hall, who also reviewed the social studies portion of the Commons, responded to plaintiffs' leading questions in a manner that leaves the record quite ambiguous. While she testified that she had in mind "the total curriculum" including "the general education [and] professional education required courses," she also had in mind "courses in the Division of Social Sciences."¹⁸² But Mrs. Hall was specifically asked whether she would have answered

"any *differently* if you were considering only . . . the social studies courses required as a part of the general education requirement, as opposed to the total curriculum required in Morris. Ly social studies majors.

"A: That is not clear to me."¹⁸³

Thus, only two of the twenty content review panelists for the Commons deposed by plaintiffs possibly misunderstood the instructions in a manner that may have affected their judgments.¹⁸⁴

¹⁸¹ Reitzer deposition, p. 27 (Ex. P-4) (Emphasis added).

¹⁸² Hall deposition, p. 80 (Ex. P-1).

¹⁸³ *Id.* at 84-85. (Emphasis added).

¹⁸⁴ Plaintiff also claims that Dr. Kimball, who reviewed the Literature and Fine Arts portion of the Commons, had only a superficial knowledge of the

The evidence for the claim that five content review panelists for the area tests made mistakes is also quite spotty.

Vinson: Plaintiff singled out Dr. Vinson as having had in mind teacher education majors who took Spanish to satisfy their language requirement, rather than those who were majoring in Spanish.¹⁸⁵ He testified that he marked items as content appropriate only if they were covered in the *texts* used in the four-semester general sequence and did not include course-related outside work as well.¹⁸⁶ Each of these decisions by Dr. Vinson could only have had the effect of *reducing* the number of items that he judged to be content appropriate.

Platt: Dr. Platt, who also reviewed the Spanish area test, is another individual who seemed quite shaken by plaintiffs' interrogation and it is difficult to tell from his deposition precisely what he had in mind.¹⁸⁷

Solomon: Dr. Solomon performed his tasks primarily with reference to social studies majors rather than social studies and history majors. But the record reflects that, in dealing with specific questions, Dr. Solomon restricted his answers to questions dealing with his field of specialization, geography.¹⁸⁸ Thus, it is entirely unclear whether his consideration of history majors in performing this task would have affected the nature of his response. Insofar as the overall test content task is concerned, it is unclear whether his judgment, taking into account both history majors and social studies majors, would have led him to state whether the test was more or less appropriate.

pertinent course requirements. But even a cursory review of Dr. Kimball's deposition indicated that his memory had faded substantially, and that, even more importantly, he was quite shaken by his deposition and by plaintiffs' leading interrogation.

¹⁸⁵ Plaintiffs' Proposed Findings of Fact, ¶ 191(c)(1).

¹⁸⁶ Vinson deposition, pp. 17-18 (Ex. P-22).

¹⁸⁷ Dr. Platt did testify that he thought that his task was to:

judge whether or not at least 90% of the students in the group . . . in the class that I am teaching, whether this material would be taught. I don't know, I don't know how else to interpret this sentence then. I, I assume that people who are going to be wanting to teach Spanish, uh, have the ability, have taken the courses to equip themselves to be teaching it. Platt deposition, p. 30 (Ex. P-24).

¹⁸⁸ Solomon deposition, pp. 21-22 (Ex. P-52).

4. The Criticisms Applicable Only to the Knowledge Estimation Panels Have No Foundation in the Record

Plaintiffs concentrate their criticisms on the Knowledge Estimation Panels because these judgments support the cut-off score recently adopted by the State Board and indicate that the previous cut-off score, which plaintiffs initially challenged, is too low. Plaintiffs advance two criticisms: the judges' alleged lack of teaching experience at the elementary or secondary school level; and the nature of the hypothetical group used.

a) Teaching experience of judges at elementary and secondary levels

Plaintiffs contend that the knowledge estimation judgment "was likely to be confounding" for those panelists who did not have direct experience as teachers in primary or secondary schools and that 16 of the 40 Knowledge Estimation Panel members had no such prior experience.¹⁸⁹ It is argued that panel members without such experience would tend to focus on the criterion "minimum knowledge to graduate" rather than the "minimum knowledge necessary to teach effectively."¹⁹⁰ As discussed above, the ETS decision not to include practicing high school and elementary school teachers on the panels is fully supportable and plaintiffs' arguments with respect to the Knowledge Estimation Panels are not persuasive.

Even assuming the accuracy of plaintiffs' numbers,¹⁹¹ there are three additional responses to plaintiffs' specific contention regarding the ability of non-teachers to apply the "teach effectively" standard properly: (1) the testimony does not support the hypothesis that those who have not taught at the elementary or secondary level cannot or did not focus properly on the amount of knowledge required to teach effectively; (2) there is no evidence that those who had such elementary or secondary teaching experience would have applied a more *lenient* standard; and (3) the reliance on only that aspect of the dual standard relating to the knowledge required to graduate from a teacher training program would in *every* instance produce the equivalent or a more *lenient* standard.¹⁹²

¹⁸⁹ Plaintiffs Proposed Findings of Fact, ¶¶ 168, 179.

¹⁹⁰ *Id.* at ¶ —.

¹⁹¹ Whereas plaintiffs assert that only 24 of the 40 members deposed had actual teaching experience, there is evidence indicating that the figure may be 28, Goldstein deposition, p. 31 (Ex. P-140), or as high as 31. Plaintiffs' summaries (Ex. P-134).

¹⁹² See discussion of effect of dual standard at pp. 91-95 below.

b) *Nature of the hypothetical group*

The members of the Knowledge Estimation Panels were given the following instructions:

Now you will be asked to draw upon your experience to construct a hypothetical group of persons, each of whom, in your judgment, has the minimum amount of academic knowledge to complete the teacher education program required for certification in South Carolina, *and* has the minimum amount of knowledge to teach effectively.¹⁹³

Plaintiffs assert that this task was complex, abstract and difficult. It is also a task that every faculty member performs, at least in part, every time he or she writes and grades examinations for students enrolled in courses at the teacher training institution. The dividing line between a minimally passing and a failing grade must be established by a faculty member in designing an examination in the subject he or she is teaching in order not to set the level of difficulty of the questions so high as to exclude the minimally qualified student from demonstrating the level of knowledge that he or she has. That dividing line must be assessed again in grading examinations and, because the delivery of a failing grade has such important consequences, most conscientious faculty members pay great attention to their conception of what the minimally qualified student should be able to do on their examination in order to achieve a passing grade. The Knowledge Estimation Panel members were asked to draw upon this knowledge in setting the dividing line between low passing and failing work and to apply their conception of the minimally qualified student to test questions written by others.

Plaintiffs apparently do not take the position that it is improper to ask a group of judges to form a hypothetical group and to make judgments based on their conception of that hypothetical group. Plaintiffs take issue only with the particular hypothetical group that was requested by the instructions.

Aside from their general argument about the alleged difficulty of the task, plaintiffs make two related arguments: first, that a dual standard was used requiring the judges to conceive of minimally qualified students both with respect to graduation from a teacher training institution *and* with respect to effective teaching; and second, that two different types of hypothetical groups were required—the persons reviewing the Commons needed to use a group of all teacher candidates and the persons reviewing the Area Examina-

¹⁹³ REPORT, p. G-3 (Ex. S-MAN-32A) (Emphasis added.)

tions needed to use a group of teacher candidates majoring in the subject covered by the test. Plaintiffs allege that the judges could not and did not comply with the instructions in either of these regards. Neither of these allegations is supported in the record to the extent claimed by plaintiffs.

(i) *the dual standard*: In theory, the nature of the approved program in South Carolina should ensure that the minimum amount of knowledge required to graduate from an approved program is *the same as* the minimum amount of knowledge required to teach effectively. However, the approved program approach has certain deficiencies that result in a variety of approaches to teacher education among the 25 teacher training institutions in South Carolina and a varied level of output with respect to minimally knowledgeable students who manage to graduate. Therefore, it is likely that in some instances the minimal amount of knowledge to graduate is less than the minimal amount of knowledge to teach effectively. In other cases, in institutions with very high standards and a highly qualified student body, the reverse may be true—the minimum amount of knowledge required to graduate may be more than the minimum amount of knowledge required to teach effectively. In order to take into account each of these possibilities, ETS devised a dual standard. That standard permits the judge to pick *either* the minimum amount of knowledge necessary to graduate *or* the minimum amount of knowledge necessary to teach effectively as the controlling standard in accordance with his or her perception of the teacher training institution program that was the basis for the judgments made.

Plaintiffs go to great lengths to fault six of the 20 Knowledge Estimation Panel members for departing from their instructions¹⁹⁴ by concentrating on one or the other of the standards. Since plaintiffs' interrogation usually did not establish whether the judge considered the standard he or she concentrated on (graduation or teaching effectively) to be the *higher* of the two parts of the dual standard, plaintiffs' effort is for naught. If the panelist did concentrate on one standard and it turned out to be the *higher* standard, then the result is the *same* as if no mistake were made; if the panelist concentrated on one part of the dual standard and it was the *lower* of the two, then the only effect on the study would be to *lower* the recommended score levels.

Further, plaintiffs' summaries of what they discovered from the judges are unreliable in two respects: first, they omit testimony

¹⁹⁴ Plaintiffs' Proposed Findings of Fact, ¶ 177.

indicating judges understood the nature of the dual standard;¹⁹⁵ and second, they erroneously summarize the judges' testimony about what was done. Plaintiffs single out panelists Arnold, Berlinghieri, Boykin, Powell, Basil and Rouse as having departed from the instructions. The record indicates otherwise, as described briefly below.

Arnold: Early in the examination, Dr. Arnold, who reviewed the Area Examination in German, testified that, as he understood it, he was to:

"picture in [my] mind . . . certain of our minimally qualified graduates in education who would be going into the teaching of secondary or elementary teaching of German."¹⁹⁶

Dr. Arnold pointed out, however, that it had by then become "difficult for me to [re]member all of the thoughts that went through my mind."¹⁹⁷ In response to a series of subsequent leading questions

¹⁹⁵ The following examples of statements, omitted from the summaries, demonstrate that the judges understood the dual requirement:

Douglass: The summaries suggest that Dr. Douglass, who was on the Knowledge Estimation Panel for the written English expression portion of the Commons, focused only on knowledge needed to "barely get through" specific courses. Plaintiffs' summaries (Ex. P-134). It omits the explanation that this was because he "would consider that the knowledge would be essential for anyone to be an effective teacher." Douglass deposition, p. 15 (Ex. P-17)

Redd: The summaries do not reflect the fact that Dr. Redd attempted to "distinguish between those persons minimally qualified to teach effectively and those persons minimally qualified to complete the academic program at The Citadel." Redd deposition, pp. 9-10 (Ex. P-11)

Waterfall: The summaries do not reflect Dr. Waterfall's statement that "[o]bviously, we were primarily concerned with those persons who are best equipped to teach, rather than just satisfying the minimum requirements for graduation." Waterfall deposition, p. 12 (Ex. P-37)

Bracey: Plaintiffs do not report Dr. Bracey's statement that "I think we are thinking about shall I say a level of . . . minimum competence you would expect these persons to possess in terms of knowledge in subject matter and ability to teach or impart knowledge to others." Bracey deposition, p. 14 (Ex. P-48)

Sherard: As a final example, one might take Dr. Sherard, who evaluated the mathematics portion of the Common Examinations. The summaries specifically refrain from setting forth the following exchanges:

"Q. All right, now, persons minimally qualified to do what?

A. To teach." Sherard deposition, p. 17 (Ex. P-30)

¹⁹⁶ Arnold deposition, pp. 7-8 (Ex. P-20).

¹⁹⁷ *Id.* at 8.

about the program at his institution, Clemson University, Dr. Arnold did note that there were very few education majors *actually* specializing in German at Clemson.¹⁹⁸ In plaintiffs' findings this latter testimony is treated as establishing that:

Panelist Arnold did not envision a group. He *tried* to envision one student meeting the instructions, he did not 'think that at that time we had one that was exactly in that category. . . .'¹⁹⁹

But plaintiffs did not altogether succeed in their efforts to confuse Dr. Arnold. Plaintiffs' findings wholly fail to apprise the court of the testimony by Dr. Arnold that appears on the very next page of his deposition:

I tried to do, uh, to envision a group of those who would be minimally qualified to, to teach, but admittedly, I feel that that was very difficult to do, but, having had some, a number of years of teaching at the secondary level, it might have been a little easier for me.²⁰⁰

It was entirely consistent with the instructions for Dr. Arnold to have pointed out that the exercise was essentially "hypothetical."

Berlinghier: Panelist Berlinghier, who reviewed the test on Chemistry, Physics and General Science, was faulted by plaintiffs for doing precisely what he was requested to do. His background was in physics, and on questions in chemistry and biology he made his estimates of how many people would know based on whether he could answer the question and whether, if he could, he regarded the question as being difficult or easy. Plaintiffs seem to suggest that Dr. Berlinghier in fact made higher estimates for subjects with which he was less familiar than he did for the physics questions.²⁰¹ In fact, just the reverse was true. Dr. Berlinghier explained that he was very low in his estimates about the proportion of minimally knowledgeable students who knew the answers to questions on chemistry, biology and geology:

¹⁹⁸ *Id.* at 10-11.

¹⁹⁹ Plaintiffs Proposed Findings of Fact, ¶ 177(e). (Emphasis added).

²⁰⁰ Arnold deposition, p. 12 (Ex. P-20).

²⁰¹ Plaintiffs cite a statement by Dr. Berlinghier that "generally, I was much more *lenient* with questions in Chemistry, Biology and Geology," thereby suggesting that his estimates of minimal knowledge in these areas were higher. Plaintiffs Proposed Findings of Fact, ¶ 177(f). (Emphasis added).

If I had no idea what the question was I put down the *lowest possible* estimate. If I knew what the answer was, but I felt it was a difficult question, I *still* put down a *fairly low* percentage. If I knew what the answer was and I felt it was general knowledge, then I put a higher percentage, but in comparing the answers to these questions, in Physics I was much more lenient.²⁰²

Thus Dr. Berlinghier followed the instructions and gave his best judgment in areas with which he was less familiar. And, contrary to the suggestion of plaintiffs' findings, he made extraordinarily low estimates for such questions, thereby reducing his own estimate of the score that would be obtained by a minimally knowledgeable teacher.²⁰³

Boykin: Plaintiffs charge Boykin with failing to follow the instructions in not distinguishing:

between the two types of minimal qualifications. He [sic]²⁰⁴ vacillated [sic] 'between what a good teacher is and what students can answer, finally choosing the percentage of students who would know.'

Dr. Boykin did not follow the instructions because she used the standard of knowledge required to graduate, and, in her view, the minimally knowledgeable graduate invariably knew *less* than the minimally knowledgeable teacher should know. In framing her answers in terms of what students actually would know, Dr. Boykin answered according to the lower of the two standards.²⁰⁵ This led to a lowering of her estimate of minimum scores.

²⁰² Berlinghier deposition, p. 16 (Ex. P-13). (Emphasis added).

²⁰³ Dr. Berlinghier specifically focused on the criterion of the minimally knowledgeable prospective teacher:

As I said before, I think I was a little bit more severe in determining the minimal knowledge for teachers as opposed to someone who is just taking the course as an elective Berlinghier deposition, pp. 13-14 (Ex. P-13)

²⁰⁴ "His" first name is Annabelle. Boykin deposition, cover page, (Ex. P-64). Plaintiffs' Proposed Findings of Fact, ¶ 177(c).

²⁰⁵ Boykin deposition, pp. 11-12 (Ex. P-64). Dr. Boykin's explanation that the minimally knowledgeable graduate would know less than the minimally knowledgeable teacher also thoroughly undermines the speculation of some of plaintiffs' experts that teachers invariably "overestimate" what their students have learned. See p. 66 *infra*.

Powell: Panelist Powell is also faulted for doing exactly what she was supposed to do. The panelists were instructed to narrow their hypothetical group to those individuals who in fact were pursuing teaching careers. Thus, there is no "mistake" reflected in plaintiffs' proposed finding that:

Panelist Powell narrowed her group to those who did serve teaching careers.²⁰⁶

(ii) *the different types of hypothetical groups:* There were two types of hypothetical groups to be used by the judges. Those who were judging the Common Examinations were asked to formulate a hypothetical group spread over all majors in a teacher training program because persons with all specialties take the Commons.²⁰⁷ The judges who were judging the Area Examinations were asked to formulate a hypothetical group composed only of majors specializing in the subject matter covered by the test (for the test in Art Education, the judges would use a hypothetical group of students majoring in art) because those would be the persons for whom the test was intended.²⁰⁸

Plaintiffs claim that some of the judges misunderstood these instructions and used the wrong type of hypothetical groups: judges on the Commons used hypothetical groups of majors, and judges in the Area Examinations used hypothetical groups of non-majors. Plaintiffs assert that at least 10 of the 20 Knowledge Estimation Panel members who reviewed the Commons made this mistake²⁰⁹ and offer their summaries of the judges' depositions as support for that proposed finding. Plaintiffs' summaries are incomplete and misleading in the following respects:

Douglass: The summaries suggest Dr. Douglass did not recall that he had reviewed a portion of the Commons.²¹⁰ However, Dr. Douglass stated in his deposition that he had been reviewing the Commons. At the outset of Dr. Douglass' deposition, the following appears:

²⁰⁶ Plaintiffs Proposed Findings of Fact, ¶ 177(d).

²⁰⁷ These instructions are set out in the REPORT at p. G-3 (Ex. S-MAN-32A).

²⁰⁸ These instructions are set out in the REPORT at p. G-5 (Ex. S-MAN-32A).

²⁰⁹ Plaintiffs' Proposed Findings of Fact, ¶ 174.

²¹⁰ Plaintiffs' summaries, Ex. P-134.

Q. And, uh, were you assigned to the Knowledge Examination Panel in the, uh, for the common exam, that portion dealing with written English and expression?

A. Uh, yes.²¹¹

Later in his deposition, in response to a series of leading questions which left the record quite confused as to the distinction between the Area Examination in English literature, that portion of the Common Examinations in literature and fine arts, and the portion of the Common Examinations on written English expression in fact reviewed by Dr. Douglass, he did indicate that he thought what he reviewed was not part of the Commons.²¹² It is on this confused portion of the record that plaintiffs' summary relies in suggesting that Dr. Douglass thought he was reviewing an area test. Further evidence of plaintiffs' erroneous reading of the transcript is found in his testimony that he thought of the course called Composition for Teachers in forming the hypothetical group.²¹³ Dr. Douglass testified that Composition for Teachers was "required of all undergraduate education majors," with the exception of those who had not performed well in Freshman competition and were required to take a special "advanced composition course."²¹⁴

Redd: The summaries also suggest Dr. Redd did not know what test he was reviewing. His deposition testimony indicates he was reviewing the Commons.²¹⁵ Dr. Redd described what he did by saying that he had asked himself whether "the weakest student in a class that I had taught [would] be able to know the answer to any given question."²¹⁶ Only in response to subsequent leading questions did he suggest that he was thinking of the specific subject matter areas—literature and the fine arts—that were covered by the test. Even then, when specifically asked whether he wasn't really thinking of a person "who would be minimally qualified to teach English

²¹¹ Douglass deposition, pp. 2-3 (Ex. P-17).

²¹² *Id.* at 8.

²¹³ *Id.* at 16-17.

²¹⁴ *Id.* at 12-13.

²¹⁵ Redd deposition, pp. 2-3 (Ex. P-11).

²¹⁶ *Id.* at 9.

effectively," he responded that he was doing so with "the fine arts and such tacked on."²¹⁷

Foster: The summary indicates that Dr. Foster, who reviewed the written English expression portion of the Commons, focused on English majors, and in response to some leading questions, she did testify to this effect.²¹⁸ But a more careful review of her deposition indicates that she had a more general group in mind as she testified:

I not only tried to hypothesize a particular group of students because communications skills cannot really be limited. . . .

* * *

. . . I didn't have to really distinguish because I thought of courses that all students were exposed to, fundamental, and, at that point, I am not concerned, I don't even know what their major is on the Freshman level, . . . if I were teaching advanced Junior and Senior level courses, I would have more or less the students who were heading for teaching positions in English. But on the Freshman and Sophomore level, when students are trying to meet basic requirements, they have to have 12 hours in English, they may be going to teach history or Physiology, I don't know, so my requirements in, uh, Freshman English wouldn't be different for anybody, the same for any student.²¹⁹

Waterfall: Dr. Waterfall also reviewed the test on written English expression. The summary strongly suggests that, in reviewing the test he was assigned, Dr. Waterfall was thinking primarily of those who would be majoring in English. A review of the deposition tells quite a different story. On direct examination, plaintiff's counsel found that Dr. Waterfall had done exactly as he had been asked.²²⁰ Dissatisfied, counsel returned to re-direct, re-re-direct and

²¹⁷ *Id.* at 10.

²¹⁸ Foster deposition, pp. 20-21 (Ex. P-34).

²¹⁹ *Id.* at 20, 23.

²²⁰ While recognizing that "nothing is infallible," Dr. Waterfall "tried to answer the questions, tried to determine a percentage as closely as I could." Waterfall deposition, p. 15 (Ex. P-37). When asked to "define for us the prospective teachers you had in mind," Dr. Waterfall stated that they were those "who were in primary and secondary education in South Carolina schools." *Id.* at 11. When asked the standard he applied in defining the minimally qualified group, he stated that:

There is a difference here. Obviously, we were primarily concerned with those persons who are best equipped to teach, rather than just satisfying the minimum requirements for graduation. *Id.* at 12.

re-re-re-direct examination of Dr. Waterfall, attempting through further leading questions to elicit testimony that would be favorable to his case. On re-redirect, plaintiff elicited the statement that Dr. Waterfall "was thinking primarily of those who would be teaching English in one way or another."²²¹ And even then, Dr. Waterfall hastened to correct any possible misunderstanding:

I think whether a person teaches math or what have you, that person should be able to express himself, that, after all, this is an influence upon the teacher, and, if a person is unable to communicate effectively, then—I feel that English is not just for people who are teaching English. Grammar is not just for those who are going to be teaching it. Every field, whether it be mathematics or what have you, a person is going to be called upon to express himself in some way. . . .²²²

On re-re-direct, plaintiff tried once more. Dr. Waterfall stated simply that:

I would say I had the person in mind who was going to go out and teach, the minimum requirement for the person who was going to go out and teach.

Q. And teach what?

A. And teach period, because this is only one area of the Teacher Examination.²²³

Paysinger: Ms. Paysinger, who served on the panel of judges for the written English expression portion of the Commons, did indicate that she had English majors in mind. But Ms. Paysinger, who among other things taught required Freshman English, thought that the test she reviewed was quite simple. Her testimony was as follows:

Q. Would you say then that a person who had completed your courses would be generally prepared to answer those questions, the sample questions, you saw on the NTE?

A. Heavens, yes. That exam is not hard. Anybody who passed my course ought to be able to answer those questions. I don't say that's going to make him a good teacher, but anybody who passed the course ought to be able to answer those questions.²²⁴

²²¹ *Id.* at 32.

²²² *Id.* at 32-3.

²²³ *Id.* at 34.

²²⁴ Paysinger deposition, p. 22 (Ex. P-36).

Tribble: Dr. Tribble, whose teaching area was economics, reviewed the Social Studies portion of the Commons. He was led by plaintiffs into testifying that, in reviewing this test, he had the minimal knowledge required of high school social science teachers in mind.²²⁵ However, he later testified that he in fact had other people in mind.²²⁶ Thus, Dr. Tribble's deposition provides little support for plaintiffs' assertion that the failure to concentrate on all education majors would have lead to an *elevation* of the knowledge estimates.

Ellis: Professor Ellis, who reviewed the literature and fine arts section of the Commons, was similar to Dr. Tribble. He did state that, in reviewing his test, he was thinking "primarily of people who would be teaching in disciplines for which they had been specifically prepared," that is literature, art and music.²²⁷ However, Dr. Ellis himself suggested that he had a broader group in mind. For example, he noted that:

It isn't infrequently that a teacher is called upon to teach a discipline outside his field of preparation and I am sure I had that in mind when I was making these evaluations but not primarily.²²⁸

Mereover, Dr. Ellis felt that a substantial portion of the questions were so general "that almost any literate person would be able to answer a considerable number."²²⁹ Indeed, he thought that a "conservative" estimate would be that 90 percent of *all* teacher education majors could answer half or more of the questions he reviewed.²³⁰

²²⁵ The following is typical:

Q. Did you limit your consideration to persons who would be teaching in the area of Social Science in the secondary school?

A. I would say yes. Tribble deposition, p. 15 (Ex. P-40).

²²⁶ On cross examination, Dr. Tribble was asked whether his answers would have varied "if you had considered teachers teaching in any field in high school, such as Math?"

A. Perhaps I'm making a mistake here in referring to high school because I did have in the back of my mind elementary school as well. If the answers would vary, they would only vary minimally. With respect to the elementary school teachers, they would not vary at all. *Id.* at 18.

²²⁷ Ellis deposition, p. 14 (Ex. P-33).

²²⁸ *Ibid.*

²²⁹ *Id.* at 21.

²³⁰ *Id.* at 24-25.

Finally, plaintiffs specifically asked Dr. Ellis whether his answers would have been any different if he had been asked to consider the amount of literature or fine arts "that persons preparing to teach in other fields would have needed to learn to complete the teacher training program." His response was that "I don't believe they would have."²³¹

Hasty: Dr. Hasty first testified that his hypothetical group was made up "principally" of his "undergraduate students in [his] introductory Political Science courses."²³² In response to a leading question about whether he could distinguish between those individuals who he knew were going to enter education, as contrasted with those who would not, he stated that he thought he might have been concentrating on individuals in his classes that "were going to teach," adding, "in my particular field."²³³ It is the latter statement that is relied on by plaintiffs to establish that Dr. Hasty actually concentrated on political science majors.

The conclusions to be drawn from the depositions of the Knowledge Estimation Panel members are that these faculty members regarded the task seriously, tried to and in most cases succeeded in using the hypothetical group that was described in the instructions, and described as best they were able what they did when they made their judgments, given the passage of time, their unfamiliarity with the deposition process, and the leading questions used by plaintiffs. None of these depositions provides any basis for disqualifying the knowledge estimation data.

5. The Data from the Judges' Depositions Generally Support the Fairness and Reliability of the ETS Validity Study

Plaintiffs used a standard outline of about 25 basic topics designed to get at possible unfairness or unreliability of the ETS validity study. The data collected from responses to 13 of those standard topics are set out in plaintiffs' summaries.²³⁴ The data with

²³¹ *Id.* at 15-16.

²³² Hasty deposition, p. 4 (Ex. P-3).

²³³ *Ibid.*

²³⁴ Plaintiffs' summaries are Exhibit P-134 and deal with testimony given by the judges in the following categories: undergraduate education, graduate education, elementary and secondary teaching experience, college teaching experience (including courses taught), observation of student teachers, preparation prior to panel participation, knowledge of curriculum, content review comparison, knowledge estimation comparison, ability to identify test examined.

respect to seven of the remaining topics were not summarized because they basically support the ETS validity study. The most important general aspects in which the data collected by plaintiffs support the fairness and reliability of the ETS study are set out in detail in Appendix C and are summarized below.

a) *Qualifications of judges*

Plaintiffs' summaries include information with respect to the qualifications of the judges and demonstrate convincingly that each person selected to participate in the study met the specifications of the study design and was, in fact, qualified to perform the tasks required by that design.

Each of the 81 judges has a bachelor's degree in a field appropriate to the subjects he or she is now teaching. Further, 74 of the 81 judges have master's degrees and 54 have PhD degrees in the fields covered by the tests they were examining.²³⁵ Each of the judges had also taught one or more courses within the past three years covering the tests they were examining. All of the judges had taught at least one such course; 79 had taught two or more such courses; 72 had taught three or more such courses; and 68 had taught four or more such courses.²³⁶ The judges' college-level teaching experience was extensive: 13 judges had two to five years' experience; 25 judges had six to ten years' experience; and 39 judges had over ten years' experience.²³⁷

As a result of this training and teaching experience, each of the judges had sufficient knowledge of the curriculum, degree and graduation requirements for the tasks required in the study.

b) *Willingness of judges to participate*

Plaintiffs attempted, but failed, to prove that the judges had been coerced by some means into participating in the study. Of the 75 judges who were asked by plaintiffs about their willingness to serve, 72 indicated that they had participated willingly.²³⁸ Some regarded it as part of their faculty duties; others regarded it as a reasonable request from their dean or department chairman. Vir-

²³⁵ Plaintiffs' summaries, column 3 (Ex. P-134).

²³⁶ *Id.*, column 1.

²³⁷ *Id.*, column 7. Information on the other four judges is unavailable.

²³⁸ Appendix C, column 3.

tually all the judges who expressed an opinion felt that the validity study was a useful undertaking, and therefore considered their participation to be a helpful contribution either to the profession or to the effort to produce qualified teachers for South Carolina.

Other independent indications of the general acceptance of the purpose and design of the study among faculty members at South Carolina teacher training institutions are the number who agreed to participate, and the number who showed up at the panel meetings. The State Department of Education was able to tap 747 educators for the panel tasks within a few weeks of being notified by ETS of its requirements. Of the educators nominated to participate, 515 were selected and 456 (or 84%) actually attended the panel sessions.

c) Absence of any outside influence prior to judges' participation

Plaintiffs conducted a searching examination of virtually every judge to find out if anyone from ETS or the State Department had contacted them in any way that might affect the substance of their judgments. Plaintiffs discovered from the 56 judges asked this question that 54 of them had received materials prepared by ETS and distributed by the State Department,²³⁹ but had had no other contact with representatives of either.²⁴⁰ The judges came to the panel meetings without any prior influence from any one involved in the study.

d) Efficient administration of the study

The responses given by the judges during these depositions indicate that the selection of the judges was made in sufficient time so that the judges could consider the instructions, and if they felt they could do so, to supplement this preparation by examination of college catalogs or discussions with colleagues.²⁴¹ Of the 63 judges who were asked, 52 indicated their selection was timely and, due to the passage of time, 10 were unclear about the dates of their

²³⁹ These materials are described in the REPORT at pp. 56-57 (Ex. S-MAN-32B) and at E-34, E-36, E-38, G-1, G-3, and G-5 (Ex. S-MAN-32A).

²⁴⁰ Two of the 56 judges were unclear about any contact.

²⁴¹ The standard question was: "Approximately how far in advance of the panel meeting were you notified of your selection?"

E. Mercer deposition, p. 9 (Ex. P-51).

selection.²⁴² The written materials containing the instructions were distributed to each judge well in advance of his participation in a panel session so that the judges could examine them carefully.²⁴³ Oral instructions were given to supplement and to clarify any ambiguity with respect to the written instructions.²⁴⁴ The depositions did not identify any systematic administrative difficulty that might have affected the results of the study.²⁴⁵

e) Clarity of instructions

The judges were asked about the instructions with respect to each of the tasks to be performed. The question was typically phrased in a straightforward fashion:

Q: . . . I will ask you whether or not the instructions you were given in paragraphs 1, 2, and 3 of E6 of SM91A were clear to you?

And the answers were equally direct. Twenty-seven of the 35 judges who were members of the Content Review Panels and who were

²⁴² Of the 80 judges who were asked, 72 remembered receiving their instructions in ample time and two were unclear. Appendix C, column 5.

²⁴³ Appendix C, column 6.

²⁴⁴ The following exchange is illustrative:

Q: . . . [D]id you receive any oral instructions prior to the panel assembly?

A: Yes.

Q: And whom did you receive these instructions?

A: I don't remember the person's name but I believe it was someone from ETS.

Q: And these instructions were the day of the meeting?

A: Yes, they were more or less of the nature—in the particular form I filled out besides this one—there was 'yes,' 'no,' 'don't know' and the type of instructions were pretty much a repeat of these overviews except don't be too hasty in putting 'don't know.'

Q: And that was?

A: E34 and E35 [referring to pages of Exhibit S-MAN-32A].

Q: You say the instructions were not to be too hasty in indicating you don't know.

A: Yes, to take your time and do a thorough job.

E. Mercer deposition, p. 12 (Ex. P-51).

²⁴⁵ Appendix C, column 7. Of the 81 judges deposed, only nine commented on any administrative difficulty of any sort.

questioned on the subject said the instructions were clear.²⁴⁶ Twenty-two of the 28 Knowledge Estimation Panel members who were asked testified that their instructions were clear.²⁴⁷ Most of those who did not find the instructions to be clear on initial inspection said that they were able to understand and perform the task required after either additional reflection or oral supplementary instructions.²⁴⁸

f) Confidence of judges in their ability to achieve fair and reliable judgments

Many of the judges were asked what level of confidence they placed in their judgments. Of the 60 who were asked, 50 indicated they had used their best professional judgment or were "very confident," "quite confident" or "reasonably confident" in their judgments.²⁴⁹ This is an overall assessment by the judges of the reasonableness of the task they were asked to perform, their qualifications to perform it, and the diligence with which they went about their work.

C. Plaintiffs' Criticisms of the Data Analysis Are Inconsequential

Plaintiffs have no serious criticism of the many intricate data processing or data analysis procedures used by ETS to convert the judgments of the 456 panel members into corresponding minimum composite score levels that could be used for making certification decisions in South Carolina. Plaintiffs point to two data analysis choices made by ETS: treatment of the more/less/same judgments of the Content Review Panels comparing the relative emphasis on various topics on the test with the relative emphasis on those topics in the curriculum;²⁵⁰ and treatment of the "do not know" responses by the Content Review Panel members in making the content appropriateness judgment on individual test questions.²⁵¹

1. Treatment of More/Less/Same Judgments

Plaintiffs fault the analysis of the data from the relative topic emphasis review performed by the Content Review Panel members

²⁴⁶ Appendix C, column 9.

²⁴⁷ Appendix C, column 10.

²⁴⁸ See discussion at p. 17 *supra*.

²⁴⁹ Appendix C, column 8.

²⁵⁰ Plaintiffs' Proposed Findings of Fact, ¶¶ 198-99.

²⁵¹ *Id.* at ¶¶ 198, 201-02.

because, in preparing an index of the composite judgments of the jury panel members, ETS permitted judgments of "more" to cancel out judgments of "less." This method of analyzing the data was used because each of the "more" and "less" judgments was considered to represent the same degree of difference, only in opposite directions.²⁵³ The resulting indices of topic difference thus reflected both the degree of difference in topic emphasis, and whether, on a composite basis, the divergence was predominantly toward "more" emphasis in the curriculum than the test or predominantly "less."²⁵³

Plaintiffs' expert, Dr. House, opined that the cancellation of mores and lesses constituted "an improper aggregation" of the data and that the only "proper summing" was to treat all judgments of either "more" or "less" as representing divergence, regardless of the direction in which they diverged. In contrast, another of plaintiffs' experts, Dr. Thorndike, observed that what Dr. House proposed was "... a possible analysis. Whether it would be better than the one that was done I am not sure."²⁵⁴ Moreover, as Dr. Linn pointed out, still a third method of dealing with the data, analyzing the "modal" or predominating of the three response possibilities, tended to confirm the accuracy of the index actually used.²⁵⁵

There are a number of ways this data could be presented. ETS chose one acceptable alternative. All of the raw data were retained and made available so that the accuracy of the ETS method could be checked and so that the State and others could perform the data analysis they thought warranted.²⁵⁶

2. Treatment of "Do Not Know" Responses

The forms for collecting responses with respect to one of the four content review judgments and the single knowledge estimation

²⁵³ REPORT, p. 78 (Ex. S-MAN-32B).

²⁵³ *Id.* Table 19, pp. 79-85. The righthand column of this Table provides an index which could range from plus one—indicating unanimous agreement that the curriculum gave more emphasis to the topic than the test—to minus one—indicating unanimous agreement that the curriculum gave less emphasis to the topic than did the test—with an index number of .000 indicating balanced agreement that curriculum emphasis were the same.

²⁵⁴ Thorndike deposition, p. 53 (Ex. P-110).

²⁵⁵ Linn deposition (Vol. II), pp. 28-31 (Ex. P-116).

²⁵⁶ Plaintiffs subpoenaed all of the raw data, which is stored on computer tapes, and used it to run a number of analyses.

judgment included a "do not know" response option.²⁵⁷ The judges on both panels were instructed:

If you feel that your experience provides you with no basis whatsoever for making a judgment about one of the questions, you may *circle* "DNK" (for "Do Not Know"). The DNK category is not to be used simply because you have difficulty in deciding [which answer to choose]; you are to make a decision even if it is a difficult one. The DNK category is to be used *only* when you have *no* basis for making any judgment.²⁵⁸

The data collected from the judges on the Content Review Panels included a relatively modest number of "do not know" responses. When analyzing the data, ETS had three choices: to count the "do not know" responses as "yes" answers, as "no" answers, or to not count them at all. The instructions were intended to encourage judges to pick a "yes" or "no" category if they had any basis at all for making the judgment, even if it was a difficult one. Those instructions necessarily produced more "no" answers than might otherwise be the case.²⁵⁹ However, ETS determined to use these instructions in order to have a sound professional basis for the treatment of the "do not know" responses.

ETS decided to treat the "do not know" responses as neither "yes" nor "no" responses, and not to include them at all in the data analysis. In so doing, ETS assumed that if the judges who selected a "do not know" response had had a basis for responding either "yes" or "no," those responses would have been distributed between "yes" and "no" in the same way that the responses of the judges

²⁵⁷ Both of these forms were used for the question-by-question review. The forms are set out at pp. E-40 and G-7 of the Appendix to the REPORT (Ex. S-MAN-32A).

²⁵⁸ REPORT, pp. E-38 (content review instructions), G-3 (knowledge estimation instructions) (Ex. S-MAN-32A).

²⁵⁹ For example, a judge who teaches chemistry is considering a question from the Area Examination in Chemistry, Physics and General Science that covers elementary geology (part of the general science requirement). The judge knows that the general science course, or the science requirement, at his institution covers geology, but has never taught that course or become familiar with the course content through other means. Therefore, he or she does not know whether the information on methods of rock formation called for by the question was available to 90% of the students. In this instance the judge might be inclined to use the "no" category, but if he or she really has no basis at all for making the decision, the "do not know" response is selected. Inspection of the curriculum at that institution is equally likely to result in a finding that rock formation is or is not covered by the science courses.

who did respond either "yes" or "no" were distributed. The raw data on "do not know" responses are available for analysis by others.

Plaintiffs' experts contended that the "do not know" response should be treated as the equivalent of "no" with respect to content appropriateness.²⁶⁰ However, their testimony reflects no underlying agreement or reason for these views. Dr. House could do no more than assert that if they said "... they didn't know, that means that they didn't agree."²⁶¹ Similarly, Dr. Bartlett thought the questions to which "... the judges responded, I don't know, are less likely to have been included in the curriculum than would be the items where they did know that they were included."²⁶² Dr. Bartlett's rationale for this assertion was even more confusing, when he conceded that his contention that "they are more like the items that they would say no, they weren't included," was such that it "may not be clear in that extreme [sic]."²⁶³

Dr. Goldstein, on the other hand, was of a contrary view. He was inclined to believe that the instructions for the "do not know" response were such as to discourage its use. In contrast with plaintiffs' other experts, he thought that the judges would resort to "do not know" only when they really had no basis for judging. And this was exactly what the instructions were designed to foster. Dr. Thorndike offered a different approach. He testified that:

I am not sure that [a "do not know" response] should be counted as a complete no. Maybe it should be counted as a fraction of a no or something of the sort but I have no real rationale for deciding how much weight to give it on the negative side. I would be *inclined* to give it some.²⁶⁴

Plaintiffs offer an alternative analysis of the data consistent with the House/Bartlett opinion that each of the "do not know" responses

²⁶⁰ They apparently advanced no similar contention with respect to the "do not know" response given by the Knowledge Estimation Panel members.

²⁶¹ House deposition, p. 52 (Ex. P-87).

²⁶² Bartlett deposition, p. 77 (Ex. P-109).

²⁶³ "Now, if a faculty member knows it is in the curriculum, through first-hand or even second-hand knowledge, he is going to know it. In courses that they taught themselves they would know what was not included. In courses that other people taught, they might say, I don't believe that Professor X includes that in his curriculum, but I am not sure, so I will say I don't know." *Ibid.*

²⁶⁴ Thorndike deposition, p. 67 (Ex. P-110) (Emphasis added).

is the functional equivalent of a "no" response.²⁶⁵ They have not offered an alternative analysis consistent with Dr. Thorndike's position that each "do not know" response be counted as 1/10 of a "no" or 1/5 of a "no." That analysis would, of course, fatally weaken their contention that more of the tests should have been put in the "special review" category by ETS.²⁶⁶

D. The ETS Study Design Was Rigorous and, if Anything, Created a More Difficult Standard for the Tests to Meet than Was Required Under the Circumstances

Because none of plaintiffs' contentions regarding the design, execution, and data analysis aspects of the study can withstand scrutiny, plaintiffs seek, in summary, to create an issue with respect to what they call "experimental bias." Conceding *arguendo* that the ETS choice with respect to the alternatives available was reasonable and professionally acceptable, plaintiffs argue that the sum of these reasonable choices was slanted in the direction of the result South Carolina wanted to reach. This argument is also fundamentally flawed. The opinion of the single expert upon which plaintiffs' argument is based is not credible; and most of the choices made by ETS among professionally acceptable alternatives created a more rigorous standard for the tests than was necessary under the circumstances.

1. The Expert Opinion on which Plaintiffs' Conclusion Is Based Should Be Viewed with Skepticism

The existence of systematic experimental bias is an issue as to which expert opinion would be relevant. Plaintiffs deposed six experts with respect to the study, Drs. Thorndike, Medley, House, Bartlett, Goldstein and Linn, and cross-examined two others, Drs.

²⁶⁵ In addition, Plaintiffs' recomputation counted *both* responses of "should not know," as well as those of "do not know," as being tantamount to "inappropriate." In fact, the responses of "should not know," which appeared on the data furnished to plaintiffs by ETS, reflected errors in the production of materials for the study that led to uninterpretable responses. REPORT, pp. 65-66 (Ex. S-MAN-91B). These "should not know" responses were left out of the ETS data analysis, and properly so. In contrast, plaintiffs counted every possible response, including these production errors, as tantamount to the response of "no."

²⁶⁶ See discussion of "special review" *supra* at p. 71. Placement of a test in the "special review" category may have no effect on the study results and plaintiffs do not contend that if their analysis had been adopted, any of the study results would have been changed.

Manning and Guion. They cite testimony from the depositions of three of those experts, Dr. Thorndike, Dr. Bartlett and Dr. House, as support for the proposition that procedures selected by ETS "had 'the effect of leading to an outcome that is consistent with the investigator's hypothesis.'" ²⁶⁷ In fact, Dr. Thorndike and Dr. Bartlett were never asked whether, in their opinion, there was systematic experimental bias in the study design. Nor were Drs. Medley, Goldstein, Linn, Manning or Guion asked that question, despite the fact that such opinions are essential to plaintiffs' contention.²⁶⁸ Only one expert, Dr. House, made such a claim.

Of all of plaintiffs' experts, Dr. House would seem by far the least qualified (if nevertheless the most willing) to make this allegation. He has shown a predisposition to be critical of many educational uses of standardized tests;²⁶⁹ he has been criticized in a distinguished educational journal for being predisposed to a particular result, rather than careful and detached, in his review of empirical data;²⁷⁰

²⁶⁷ Plaintiffs' Proposed Findings of Fact, ¶ 198.

²⁶⁸ Despite plaintiffs' suggestion from the order in which the proposed findings are set out that Dr. Linn had such an opinion, the record is entirely to the contrary. Dr. Linn, an outside ETS reviewer who was called as a witness by plaintiffs, made quite clear that he did not regard the study as "biased." Indeed, he regarded the proposal as being of "excellent technical quality overall." Letter from Linn to Stewart, dated Sept. 27, 1975 (Ex. P-MAN-18); Linn deposition (Vol. I), p. 146 (Ex. P-116).

²⁶⁹ Dr. House evidently regards the use of standardized tests in educational "accountability programs" that measure student learning gains and teacher performance as leading to "misshapen lives" and as being part of "schemes" that are "simplistic, unworkable, contrary to empirical findings, and ultimately immoral." House, *The Price of Productivity: Who Pays?* in *TODAY'S EDUCATION* 65-66, (published by plaintiff NEA) (Sept.-Oct. 1973).

²⁷⁰ Dr. House was qualified by the plaintiffs to testify to the "experimental bias" in the ETS study by virtue of his special expertise in "the politics of evaluation, that is how studies are affected by the circumstances in which they occur. . ." House deposition, p. 25 (Ex. P-87). This subject is one on which Dr. House has published a book entitled *THE POLITICS OF EDUCATIONAL INNOVATION*. A reviewer of that book, while finding it intriguing, had the following to say:

"The book neither organizes existing research nor offers much new empirical data. In fact, one might apply to this book a caution House urges for interpreting the rhetoric of innovators: such rhetoric is "best understood as a statement of ideology . . . but it should not be taken too literally as the goals of the project by which it must be evaluated" (p. 195). In recommending innovation in federal policy, House frequently uses the researcher's style of presentation and implicitly promises balanced conclusions regarding research on innovation. In fact, he marshals only the evidence that best supports his arguments." Hutchins, *Book Review: The politics of Educational Innovation*, 46 *HARVARD EDUCATION REVIEW* 134-35 (1974) (Emphasis added).

and he has previously undertaken to advance substantially the same position as a contractor to plaintiff NEA.²⁷¹ For all of these reasons, Dr. House's qualifications to claim that the study was biased must be given the most careful scrutiny.

2. *The Choices Made by ETS Created a Rigorous Standard*

Plaintiffs cite three choices that ETS made among professionally acceptable alternatives: the majority-rule standard; the exclusion of "do not know" responses; and the treatment of the data from the more/less/same judgments of the Content Review Panels.²⁷² They argue that each of these choices was made in a way that, even if professionally acceptable, "favors" the findings of the study that: (1) the tests were highly congruent with the South Carolina curriculum; and (2) appropriate cut-off scores for each of the tests were significantly above the previous certification level of 975.

Plaintiffs' analysis on this point is easily rebutted. An analysis of all of the choices made by ETS indicates that even if plaintiffs are right about the three they have picked out, which ETS contends they are not, these three are more than balanced by choices that "favor" the expression of a wide variety of views and that have an effect that could not be predicted, before the data was received, to be favorable or unfavorable to South Carolina's position.

²⁷¹ Plaintiffs offer Dr. House "as an expert in the area of educational evaluation." Plaintiffs Proposed Findings of Facts, ¶ 167 & n. 46. A principal concern of education "evaluators" is with "accountability" programs by which educational systems seek to ascertain the extent to which teachers contribute to the successful performance of their students. Such "accountability" programs typically make use of standardized tests. NEA has been extremely critical of such teacher "accountability" programs, and has staunchly resisted their implementation. *E.g.*, Bain, then President-Elect of NEA, *Self-Governance Must Come First, Then Accountability*, PHI DELTA KAPPAN 413 (April 1970); *See also Hearings before the Senate Subcommittee on Education on S. 3883 and S. 4167*, 91st Cong. 2nd Sess. at 122-23 (1970).

Several years ago, after the Michigan Department of Education had instituted a statewide accountability system, using standardized tests to measure pupil performance, NEA (and its local affiliate, the Michigan Education Association) sought to "evaluate" this program. They hired Dr. House and several of his colleagues to undertake this task. House deposition p. 94 (Ex. P-87).

²⁷² ETS' reasons for establishing the majority rule or 51% standard are set out at pp. 67-72 above; the rationale of excluding the "do not know" responses is set out at pp. 105-08 above; and the basis for ETS' judgment with respect to the more/less/same data analysis is set out at pp. 104-05 above.

1. Use of a 90% standard for the knowledge estimation task. This standard—that 90% of all students would have the opportunity to learn the subject matter of the question—is so high that it calls for virtually complete coverage of the student body of the teacher training program. A much lower standard would have been professionally acceptable.²⁷³ The 51% standard that is so criticized by plaintiffs operates in every case in conjunction with the very rigorous 90% standard²⁷⁴ and thus any leniency of the 51% standard is fully balanced.
2. The judgments were drawn from two different types of panels. This minimized the possibility of confusion or misconception about the tasks,²⁷⁵ but also increased greatly the number of faculty members participating in the study. Using 456 judges instead of some much smaller number, which would have been professionally acceptable, increased the possibility of diversity of judgments and the likely range over which those judgments would extend.
3. Two separate sets of panels were created for comparable tests in the Common Examinations and Area Examinations. All of the tests included in the Common Examinations except Test A (Professional Education) have counterparts among the Area Examinations. One Content Review Panel and one Knowledge Estimation Panel could have been used to review each of these tests (for example, the social studies part of the Commons would be reviewed by the same panel as reviewed the Area Examination in Social Studies.) ETS chose to create duplicate panels to minimize the possibility of confusion about the reference groups (all students in the case of the Commons and students majoring in the subject in the case of the Area Examination). This choice contributed to increasing the number

²⁷³ One of plaintiffs' experts suggested an 80% or 85% standard. House deposition, p. 60 (Ex. P-87).

²⁷⁴ Questions were classified as content appropriate only if 51% of the judges found that 90% of the students would have had an opportunity to learn the subject matter. If the study design had called for a standard lower than 90%, a standard higher than 51% might have also been used.

²⁷⁵ REPORT, p. 20 (Ex. S-MAN-32B).

of participants in the study and the likelihood of a wider range of responses. There was no way to predict in advance that this decision would "favor" one result or another.

4. All 25 teacher training institutions in the State were included in the study rather than just a sample. This increased the size and complexity of the study and provided the maximum chance for divergent views on the appropriateness of the tests.
5. Institutions with predominantly black student bodies were over-represented on the panels. These institutions contributed only 18% of the graduates, but provided 21% of the nominees and 23% of the participants.²⁷⁶ The view that judges from black institutions would have systematically different judgments than those from white institutions was given a more than fair test and if that view had been correct, this was a choice that increased the possibility of a finding that the tests were not congruent with the curricula.
6. Small institutions were overrepresented on the panels. The 12 smallest institutions in the state produced only 19% of the graduates of teacher training programs, but contributed 33% of the total faculty nominees.²⁷⁷ Smaller colleges are more likely to have specialized programs or truncated curricula. Selecting a larger proportion of the judges from these institutions increases the likelihood of a finding that the tests are not fairly matched with the curricula.
7. The faculty members from black institutions were spread out among all the panels in a prespecified form in order, whenever possible, to ensure that each panel included at least two such representatives.²⁷⁸ This ensured that the input of black institutions would be reflected throughout the study and maximized the contribution of that input. If black institutions had had a decidedly different view of the appropriate curriculum for teacher training, the impact of that view would have been felt throughout the study.

²⁷⁶ REPORT, pp. 36, 38-41 (Ex. S-MAN-32B).

²⁷⁷ *Id.* at p. 29.

²⁷⁸ *Id.* at 35.

8. The study included a method of data analysis that permitted consideration of the consistency of the panel judgments. In 34 of 52 cases, the panels of judges were divided into half panels. Subject matter coverage, institutional size and student body racial composition were taken into account in forming the half panels. These data on the consistency of the panel judgments were an additional measure of the ability of the judges to perform the tasks assigned to them and the direction in which that measure would point could not be predicted in advance.
9. The study design included a training exercise of Knowledge Estimation Panel members. A number of illustrative NTE test questions were displayed and the judges were asked to estimate the percentage of students in each quintile who would know the answer. After the judges made their estimates, they were shown the percentage of examinees in a sample that was similar to the group of South Carolina seniors who take the test for the first time (by quintile) who actually knew the answer.²⁷⁹ Only after this training exercise did the judges turn to the knowledge estimation task with respect to the study. The training exercise was designed specifically to counteract any tendency to be too stringent in judging the level of knowledge of a minimally qualified student that would have raised the score levels.
10. The response form for the Knowledge Estimation Panels included seven possible numerical responses—2%, 10%, 25%, 40%, 60%, 80% and 98%—which were weighted toward levels under 50%. Rather than being centered about the middle difficulty point of 50%, the scale is centered about 40%, a point that approximates the percentage of NTE questions to which the average South Carolina examinee was estimated to know the answers.²⁸⁰ This response scale encouraged conservative responses and was less likely to be favorable to higher score levels than other response scales that might have been chosen.
11. Panel members made individual judgments that were not influenced by the consensus of their colleagues. Although the

²⁷⁹ For example, 5% of the students in the lower 20% knew the answer, 15% of the students in the second lower 20% knew the answer, etc.

²⁸⁰ REPORT, p. 52 (Ex. S-MAN-32B).

members of the panels met as part of a large group to make their judgments, the panels were constituted solely for purposes of data analysis. The members of the panels never met as a group and never discussed their opinions. Thus members who had views that tended to make the content congruence lower and score levels higher had no impact on their colleagues who had different views.

12. Panel members were encouraged to prepare if they thought that to be necessary. Such preparation would uncover any divergence between the content of the tests and the content of the teacher training program.
13. The content review judgments were solicited in not one but four different forms. The use of four different ways for each judge to express an opinion with respect to the congruence between the content of the test and the content of the curriculum maximized the likelihood that judgments of divergence between the test and the curriculum would be given. All four forms of these judgments were considered by ETS in evaluating the level of congruence.
14. Questions from two different editions of each test were used so that an additional check for consistency could be constructed. The percentage of content appropriate questions on each of the two test editions could be compared. If a difference emerged, this would be additional ground for questioning the appropriateness of the test.
15. When the Knowledge Estimation Panel results were processed, mean percentages were computed.²⁸¹ If the median had been used (an equally acceptable procedure), very discrepant responses (those on the far ends of the scale) would have had less effect than if the mean were used. The mean gives more weight, for example, to very low estimates, than would the median. Therefore, using the mean can lower the pooled estimate of the percentage of minimally qualified students who would know the answer and thereby lower the resulting score level.

In summary, plaintiffs' allegations of "experimental bias" have no basis in either credible expert opinion or analysis of the choices with respect to the study design, execution or data analyses that

²⁸¹ REPORT, p. 104 (Ex. S-MAN-32B).

were made by ETS. The study employed rigorous standards that alone or collectively did not "favor" any possible outcome that could have been postulated at the outset of ETS' work. The study was a comprehensive, innovative, useful technique devised by ETS to meet the special legal and practical burdens faced by South Carolina, and the results of the study provide a rational basis for the State's new certification system.

CONCLUSION

For the reasons stated above, Educational Testing Service, as *amicus curiae*, respectfully suggests that this court decline to grant injunctive relief or that, if granted, any injunctive relief be structured to permit implementation of the new certification system adopted by defendant State Board of Education in June 1976, or modifications of that system that have a similar rational basis.

Respectfully submitted,

HOWARD P. WILLENS
DEANNE C. SIEMER
THEODORE S. SIMS

1666 K Street, N.W.
Washington, D.C. 20006

*Attorneys for Amicus Curiae
Educational Testing Service*

Of Counsel:

WYCHE, BURGESS, FREEMAN
& PARHAM
44 East Camperdown Way
Greenville, South Carolina 29603

JOHN S. KRAMER
General Counsel
Educational Testing Service
Princeton, New Jersey 08540

WILMER, CUTLER & PICKERING
1666 K Street, N.W.
Washington, D.C. 20006

July, 1976

APPENDICES

APPENDIX A

**ANALYSIS OF FACTS AND CONTENTIONS WITH
RESPECT TO THE COMMON EXAMINATIONS**

APPENDIX A

TABLE OF CONTENTS

	Page
A. Statement of Facts With Respect to the Underlying Validity of the Common Examinations	A-1
1. Purpose and Design	A-1
2. Method of Construction	A-4
3. Maintaining the Tests in Up-To-Date Condition.....	A-9
4. Computation of Scores	A-15
5. Properties of the Tests	A-17
a) Standard error of measurement	A-17
b) Reliability	A-18
c) Speededness	A-19
d) Validity	A-19
6. Use	A-22
B. Rebuttal of Plaintiffs' Contentions with Respect to Disqualification of the Common Examinations	A-23
1. The Tests Are Not Racially or Culturally Biased	A-23
a) Statistical compilations of scores	A-23
b) The Spencer study	A-25
c) The Medley study	A-29
d) The Cohen opinion	A-32
2. The Tests Have a Legitimate Objective	A-33
3. The Tests Are Up-To-Date	A-36
4. The Tests Are Fairly Representative	A-37

A-1

APPENDIX A

ANALYSIS OF FACTS AND CONTENTIONS WITH RESPECT TO THE COMMON EXAMINATIONS

This appendix sets out the facts with respect to the purpose, construction, scoring, and statistical properties of the Common Examinations.¹ These facts are substantially the same as those before the court in *United States v. North Carolina*, 400 F. Supp. 343 (E.D.N.C. 1975), and they support defendants' proposed findings of fact with respect to the NTE.² This appendix also summarizes ETS' rebuttal to the arguments advanced by plaintiffs' experts that the Commons are inherently inappropriate for use in the certification of teachers—entirely apart from the issues raised with respect to the validity study. The claims advanced in the testimony of plaintiffs' experts are substantially the same as those in the *North Carolina* case. 400 F. Supp. at 350.

A. Statement of Facts With Respect to the Underlying Validity of the Common Examinations

This section provides a brief review of the evidence with respect to the purpose, method of construction, scoring, statistical properties and uses of the Common Examinations, in order to demonstrate the validity of these tests with respect to their basic objective. This underlying validity of the Common Examinations supports their use in South Carolina and elsewhere in accordance with ETS Guidelines.

1. Purpose and Design

The Common Examinations were developed to provide a uniform measure of the level of achievement in professional and general education of persons completing teacher training programs. Professional education is generally defined to include psychological

¹ The 27 Area Examinations are similar to the Commons with respect to the technical aspects of construction, scoring, and statistical properties. These tests cover the subject matter knowledge that prospective teachers are expected to teach. Art teachers take the Area test in Art Education, French teachers take the Area Examination in French, and so on. These Examinations are not discussed in detail here because the plaintiffs did not make any serious effort during the discovery period either to question the underlying premises of the Area Examinations (e.g., that a French teacher should know French) or to challenge their appropriateness.

² Defendants' Proposed Findings of Fact, Section III, pp. 15-36.

and social foundations of education, and generally applicable methods and principles of teaching. General education usually includes a familiarity with the basic elements of the liberal arts: English, social studies, literature, the fine arts, science and mathematics. Both professional and general education objectives require the mastery of subjects other than those that persons completing the teacher training program will ultimately teach to their students (e.g., business education, elementary school reading, Spanish, agriculture, etc.)

The need for a uniform measure of achievement arises out of the increasing number and mobility of teachers and the wide variation in quality of teacher training programs. A primary measure of academic achievement is the grade point average achieved at the completion of the required curriculum. But teacher training institutions differ in the standards they apply to academic achievement, the content of the courses that are taught, the qualifications of faculty members and the relative quality of the student body. This is particularly true in the areas of professional and general education. The Common Examinations provide a common denominator by which candidates with different preparation and different credentials can be compared.

ETS' development of the Commons is based on three fundamental premises: (1) that knowledge of ideas, theories, principles, methods and facts other than the specific subjects to be taught are necessary to the success of the teaching effort (for example, knowledge of the psychological principles involved in learning is helpful in the teaching of mathematics); (2) that this knowledge can be systematized into academic courses and learned in the course of academic training; and (3) that teacher training institutions throughout the United States can present suitable courses for preparation of elementary and secondary teachers. Plaintiffs have not made any serious effort to challenge those basic premises.³

³ Critics in and outside the education profession have attacked these premises arguing that subject matter knowledge alone is sufficient for effective teaching (for example, all a mathematics teacher needs to know is mathematics); or that the knowledge really needed to teach effectively is how to get along with or relate to other human beings and that cannot be taught in an academic setting; or that teacher training institutions are hopelessly outmoded. There are oblique references to these sorts of contentions in the record, but plaintiffs have not undertaken to prove any of these propositions.

A-3

The Common Examinations consist of four separate tests: one test in professional education and three tests in general education.

Test A is the test in professional education. It covers the materials included in basic courses in psychological foundations of education, typically general or introductory psychology, child psychology, adolescent psychology, educational psychology, human growth and development, and testing and measurement. Test A also covers materials in social foundations of education found in courses in introduction to education, philosophy of education, history of education, the school in American society, and foundations of education. The third type of material covered by Test A is that devoted to principles and methods of teaching, such as general concepts and procedures relating to classroom management and organization, instructional behavior of teachers, the organization of instructional materials and the evaluation of classroom procedures. Test A has 110 questions (of a total of 270) and is allotted 90 minutes (out of $3\frac{1}{4}$ hours).

Test B covers written English expression. It provides a measure of the candidate's ability to use the English language correctly, and general verbal ability. The questions deal with grammatical construction, sentence construction, punctuation, capitalization and word usage. Test B has 45 questions and is allotted 30 minutes.

Test C has two subparts: social studies, and literature and fine arts. The social studies portion covers general knowledge of sociology, geography, political science, economics, anthropology and history. The literature section covers basic world literature classics and the fine arts section covers art, music and architecture. Test C includes 65 questions and has a time limit of 40 minutes.

Test D also has two sub-parts: science and mathematics. The science portion covers elementary principles of biology, chemistry and physics. The mathematics section covers basic addition, subtraction, multiplication, division, and elementary algebra and geometry. Test D has 50 questions and is allotted 35 minutes.

Each of these tests is an achievement test, which measures knowledge attained or a level of preparation achieved, rather than an aptitude test, which measures capability to attain knowledge or to achieve a given level of preparation.

Ancillary aspects of ETS' purpose with respect to the Common Examinations are to offer a test that is useful nationwide, that is secure, and that can be used over an extended period of time.

ETS endeavors to make the test useful nationwide by providing national norms. Samples are selected of the test scores of persons from the various regions of the country who take the Commons. These samples are then weighted by the number of graduates of teacher training programs from each of the regions, as compiled by the National Education Association. A region that contributes relatively more graduates of teacher training programs will also contribute relatively more to the norm. A score obtained by any candidate anywhere in the country can be compared with the norms published by ETS, representing the scores achieved by all candidates who are recent graduates.⁴

The tests are kept secure so that no candidate knows what the questions will be and no part of the results reflect an advantage in preparation for specific questions.

Scores from administrations of the Common Examinations given in any particular year can be compared with scores from administrations in prior or subsequent years because each form of the Commons is equated with each other form and the scores are expressed on the same scale from year to year. This equating and scaling is accomplished through test construction and statistical analysis processes described below at pp. A-8, A-16.

2. Method of Construction

In their present general form, the Common Examinations date back to 1961, when ETS convened a Committee of Examiners to assist it in a comprehensive re-examination of appropriate fields and subjects within fields to be covered by the Commons. That Committee consisted of seven experts with a wide range of experience in elementary and secondary education drawn from all parts of the country. These experts included present and former teachers, curriculum planners, elementary and secondary school administrators, and faculty members and curriculum experts from teacher training institutions.⁵ The Committee had two tasks: to consider in

⁴ Manning deposition, pp. 359-64 (Ex. S-8). Approximately 20% of the group on which the national norms are based are candidates who are black or members of other minority groups. *Id.* at 362. See also NTE TECHNICAL HANDBOOK 17 (Ex. S-MAN-9).

⁵ These persons were nominated by the National Commission on Teacher Education and Professional Standards, an affiliate of plaintiff National Education Association. Manning deposition, written questions on re-direct examination, Q. 32, written questions on re-cross examination (Set. I), Q. 4 (Ex. S-8).

depth the scope of the professional education tests;⁶ and to review, in more general terms, the requirements of the general education tests.

ETS staff experts conducted a research program to assist in defining appropriate subject matter for the professional education tests. This included an extensive study of the professional education course requirements of 248 teacher training institutions in 46 states.⁷ Information was gathered from each institution as to the required professional education courses, the basic textbooks used, and the outlines or syllabi of materials covered in these courses. The study confirmed that most professional education requirements included courses in psychological and social foundations of education. The study identified the psychology courses most widely required for preparation of both elementary and secondary teachers, and also found that there were common requirements with respect to courses on the social foundations of education. During the course of the study, basic textbooks currently used in teacher training programs were identified and examined, course outlines were compared, and the results of these investigations were discussed with the outside experts.⁸

There were also several important outside research studies with respect to teacher education available at this time that were reviewed with respect to the specifications of the Commons. A widely recognized study by James B. Conant examined in detail the academic preparation and certification of teachers in the 16 most populous states.⁹ Conant visited 77 teacher training institutions and his

⁶ Separate tests were under consideration at this time: psychological foundations; societal foundations; and teaching methods. These were consolidated into a single test in 1966. See pp. A-9, A-10 *infra*.

⁷ Humphry, *A Survey of Professional Offerings in NCATE-Accredited Institutions*, JOURNAL OF TEACHER EDUCATION 406 (Dec. 1963). That article is Exhibit S-MAN-39. Dr. Humphrey is a Senior Examiner in the Higher Education and Career Programs Division at ETS. Dr. Humphrey received her PhD training in educational psychology, taught for five years at the elementary and secondary school level and for one year at the college level, has extensive experience in editing textbooks for use at various levels, and has worked in the test development field for the past 15 years. Dr. Humphrey is the author of a number of professional papers and other publications. (Ex. S-MAN-40).

⁸ Manning deposition, written questions on re-direct examination, Q. 35 (Ex. S-8).

⁹ J. Conant, *THE EDUCATION OF AMERICAN TEACHERS* (1963).

319-page book describes academic requirements in both professional education and general education and the supporting rationale for these courses. A second study by James D. Koerner, another well-known researcher, used case studies of 32 teacher training institutions to analyze the quality of students and the quantity of education courses in their academic programs.¹⁰ One part of the study specifically analyzed the importance of the general education requirement. The transcripts of 435 prospective elementary school teachers and 446 prospective secondary school teachers were combined into an academic profile showing a mean of 72 semester hours in general education for elementary teachers and 96 semester hours for secondary teachers.¹¹

ETS had at hand up-to-date reviews of each of the fields covered by the general education tests because the NTE Program includes Area Examinations in each of these fields.¹² Committees of Examiners for the various Area Examinations meet regularly to review developments in these fields and large panels of outside experts create pools of questions that can be used or adapted for the general education tests. The principal task for the Committee reviewing the Commons was to consider the level of familiarity with these liberal arts fields to be required of prospective teachers and to review questions selected from the pool of available materials

¹⁰ J. Koerner, *THE MISEDUCATION OF TEACHERS* (1963).

¹¹ *Id.* at 128. This work is cited for comparison purposes by Sherwin in the 1974 study that is Exhibit P-MAN-12. For other important reference works used at that time, see Manning deposition, written questions on re-cross examination, Q. 27(c) (Ex. S-8).

¹² The Area Examinations that correspond to the general education tests on the Common Examinations are as follows:

<u>Common Examinations</u>	<u>Area Examinations</u>
Test B (Written English expression)	English language and literature
Test C (Social studies, literature and fine arts)	English language and literature Social Studies Art Education Music Education
Test C (Science and mathematics)	Biology and General Science Chemistry, Physics and General Science Mathematics.

Manning deposition, written questions on re-cross examination, Q. 29(e) (Ex. S-8).

to ensure that they reflected accurately the desired coverage of these fields.¹³

During 1961 and 1962, the outside experts and ETS staff experts worked on a series of test matrices for the Common Examinations. A test matrix indicates the required subject matter content to be covered, and the required cognitive processes to be tested (such as recall of facts, analysis of given facts, and reasoning by analogy). The matrix then becomes a guide for the questions to be included in the test by defining both the subject matter content and the cognitive process to be called for in answering the question.¹⁴ A test matrix was created for each of the tests in professional education and general education that were to make up the Commons. Further specifications were created as to the relative emphasis to be given various subject matter fields (how many questions on each subject to be included in the test) and the relative difficulty of the questions in each field.

In 1963, the experts constructed two forms of each of the tests from the specifications. The test construction process used in creating this version of the Commons has been refined by ETS through many years of experience with a wide variety of tests in a large number of professional and other fields. This process includes a number of discrete steps:

Question writing. ETS staff experts and outside experts in specialized subject matter fields write questions to fit various parts of the matrix. Specific instructions are provided for question writers with respect to the number of alternative responses, general level of difficulty, general vocabulary requirements, the framing of questions and answers and other technical matters. Each question has five alternative answers, four incorrect and one correct. Each expert generally writes from 10 to 20 questions within a particular field.¹⁵ An oversupply of questions is required at this point so that

¹³ Manning deposition, written questions on re-direct examination, Q. 35 (Ex. S-8).

¹⁴ Manning deposition, pp. 409-10 (Ex. S-8).

¹⁵ A large number of the questions on the Commons are contributed by outside experts. For example, of the 148 newly written questions for the forms of the professional education test (Test A) in use during 1975, 139 (or about 94%) were written by independent outside experts. Of the 41 newly written questions for the forms of the social studies portion of Test C in use during 1975, 35 (or about 85%) were written by outsiders. Manning deposition, written questions on re-direct examination, Q. 42-43, 106 (Ex. S-8).

replacements will be available if some questions are later disqualified.

Question screening. Each question written by ETS staff or an outside expert is reviewed independently by two ETS test specialists. This review considers whether the question meets ETS standards as to form,¹⁶ and whether the question fits the designated part of the matrix for which it was written. The newly written questions are also submitted to the outside experts for review. They examine the questions for ambiguities in language or usage and mark the correct answers. These experts also provide an independent assessment of the level of difficulty of each question. The ETS test development staff experts coordinate the responses from outside experts and weed out or revise any questions as to which the experts disagree.

Selection of questions used for equating. At least 20 percent of the questions on any new test are taken verbatim from questions on a predecessor form of that test to establish the necessary comparability of scores from one test to another over time.¹⁷ ETS staff members review previous tests and select the required number of equating questions in line with specified content and statistical requirements.

Test assembly. The newly written questions that remain after the screening process and the equating questions are assembled into the final form of the examination. Instructions are prepared, stand-

¹⁶ There must be a single unambiguously correct answer to the question; the question must be expressed in a clear, straight-forward manner and in proper grammatical form; there must be no extraneous cues about the question that might lead a non-knowledgeable candidate to discover the correct answer; and the question must meet ETS conventions with respect to style and format. Manning deposition, pp. 77-78 (Ex. S-8).

¹⁷ An established procedure is followed to insure that different tests are comparable, that is, that they represent the same level of difficulty. A designated number of questions are selected from an old form in a manner that reflects statistically the composition of that test. The designated questions are then embedded throughout the new form in approximately the same position that they occupied in the old form. When the new form is administered, performance data on the designated questions are computed and compared with similar performance data on these questions on the old form. Based on an analysis of the common questions, scores obtained on the new form can then be adjusted in a manner that reflects the level of difficulty of the new form.

The equating procedure and formulae are set out in Manning deposition, written questions on cross examination, Q. 7-13 (Ex. S-8).

ard procedures are followed to put the questions in an appropriate order throughout the test,¹⁸ and the test is printed. Proofs are examined carefully for typographical and editorial errors.

3. Maintaining the Tests in Up-to-Date Condition

After the review and development of specifications for the Common Examinations was completed in 1962, and new forms of each of the tests were completed in 1963, ETS undertook to keep the Examinations up-to-date primarily through an in-house effort.¹⁹

A team of test development specialists was assigned to the program, including experts in each of the fields covered by the professional and general education tests.²⁰ They review college textbooks used in the field, make curriculum surveys, and review the professional literature to keep up-to-date in the field.²¹ New forms of the tests were created each year from 1970 to the present.²² The test development specialists assigned to the professional education tests were assisted in developing these new forms by panels of outside experts who wrote and reviewed questions in their fields of specialty. The test development specialists assigned to the general education tests drew on the panels of outside experts assembled each year to write and review questions for the corresponding Area Examinations.

The specifications for the Common Examinations were also revised periodically. In 1966 the three separate tests in professional education were combined to provide more flexibility in designing questions and to reflect the recognition in the profession that the

¹⁸ The questions are often alternated according to content and ordered by difficulty level so that easy, moderately difficult and more difficult questions occur in series throughout the test. This serves candidate motivation and test reliability purposes.

¹⁹ Manning deposition, written questions on re-direct examination, Q. 85; written questions in re-cross examination, Q. 29(c) (Ex. S-8).

²⁰ During 1975 there were 12 ETS test development experts assigned to work on the Commons. Manning deposition, written questions on re-direct examination, Q. 116, written questions on re-cross examination, Q. 29(e) (Ex. S-8). Their qualifications are set out in Exhibit S-MAN-33.

²¹ Manning deposition, written questions on re-direct examination, Q. 96 (Ex. S-8).

²² *Id.* (Set II), Q. 5-6.

knowledge and skills acquired in these academic disciplines were used most usefully in an interrelated fashion.²³

In 1969 the specifications were revised again to reflect a growing concern in the profession that the materials include a fair representation of minority group presence and experience. A committee of minority group educators and educational administrators was assembled to assist ETS in this revision.²⁴ The Committee met at the ETS facility in Princeton with 16 ETS staff members, including every test development specialist, program staff member, supervisor, and officer with responsibility for the NTE Program.²⁵

After extensive discussion and deliberation, the Committee reached a consensus on one recommendation with respect to the professional education test and six recommendations with respect to the three general education tests. They recommended that the professional education test include more materials that reflect knowledge of minority experience and culture. They made the same recommendation with respect to the general education tests, and in addition recommended that the settings of more of the language items reflect situations that would be particularly familiar to minority candidates, that more questions be included on contemporary social developments, that more time be allotted to the written English expression test, that a type of item called sentence correction replace the type of item called error recognition, and that the social studies items be separated from the other items on Test C.²⁶

Each of these recommendations was implemented by ETS in 1970, and these changes are reflected in the example of the Common Examinations that is included in the record as Exhibit S-MAN-27. The professional education test contains 12 questions out of a total of 110 (approximately 10%) that reflect knowledge of minority

²³ Manning deposition, written questions on re-cross examination, Q. 29(a) (Ex. S-8).

²⁴ Exhibit S-MAN-31 identifies the members of this panel. Manning deposition, pp. 245, 381-82 (Ex. S-8).

²⁵ Manning deposition, written questions on re-cross examination, Q. 20(a) (Ex. S-8).

²⁶ *Id.* at Q. 20(c). The Committee also recommended that a new test related to teaching in urban schools be developed. A short time later the new Area Examination in Education in an Urban Setting was produced.

A-11

experience and culture.²⁷ Test B, on written English expression, contains seven such questions out of a total of 45 (approximately 15%).²⁸ Test C, on social studies, literature and the fine arts, contains 10 such questions out of a total of 62 (approximately 15%).²⁹

The amount of time for the written English expression test was increased from 25 to 30 minutes. This test includes only 45 questions and is not speeded, so that nearly every candidate reaches or finishes the last question.³⁰ Twenty-five of the 45 questions on this test are sentence correction questions in which a sentence is given with four parts underlined and the candidate must identify which of the underlined parts represents a grammatical error, or choose "no error," the fifth alternative.³¹

Test C was modified to separate the social studies questions from the literature and fine arts questions. The social studies portion of the test is now separately labelled and is given a separate suggested allocation of time.³²

After the review in 1969 by the special panel of outside minority group experts, ETS took the following steps to build into the test construction process a continuing review to ensure that there are no questions that on their face might be biased and that there are no question types that unfairly penalize minority group candidates who possess the requisite knowledge: (1) the panels of question writers who create and review proposed questions for the Common Ex-

²⁷ Common Examinations, Test A (Form XNTI), Questions 2, 19, 32, 37, 38, 46, 66, 76, 89, 96, 103, 106 (Ex. S-MAN-27); Manning deposition, p. 238 (Ex. S-8).

²⁸ Common Examinations, Test B (Form XNTI), Questions 122, 125, 127, 129, 135, 139, 150 (Ex. S-MAN-27); Manning deposition, p. 238 (Ex. S-8).

²⁹ Common Examinations, Test C (Form XNTI), Questions 160, 164, 165, 170, 175, 178, 188, 197, 198, 200 (Ex. S-MAN-27); Manning deposition, p. 238 (Ex. S-8).

³⁰ NTE TECHNICAL HANDBOOK, 27, 29 (Ex. S-MAN-9). Exhibit S-MAN-8 includes the test analyses performed by ETS after each new form is administered. These analyses give information with respect to the statistical properties of the tests and describe the speededness of each test.

³¹ Test B, Questions 111-35 (Ex. S-MAN-27).

³² Test C, p. 23 (Ex. S-MAN-27).

aminations include minority group members;³³ (2) minority group persons who are test specialists or subject matter experts on the ETS staff review questions prior to inclusion in the Examinations;³⁴ (3) ETS brings minority group persons who are experts in testing or related fields to ETS for sabbaticals, summer work, or work in connection with doctoral dissertations, and these persons review question types and proposed questions;³⁵ and (4) ETS has devoted substantial resources to research on testing with respect to minority group persons and since 1965 has produced 184 major studies on various aspects of these issues.³⁶

As a result, there are no questions on the Common Examinations that contain material that is offensive to any minority group or cultural group in its wording. There are no questions that are systematically answered incorrectly by minority group persons by virtue of culture-specific usage, that is, where the words used commonly have different meanings for blacks and whites. A representative proportion (10% to 15%) of the questions on each of the tests reflect the respect for and recognition of minority group experience and culture or call for knowledge relevant to minority group problems. No question type is significantly harder for minority group persons who possess the requisite knowledge than for majority group persons. No time limitation is imposed that discriminates unfairly against minority group persons. Any advantage asso-

³³ Manning deposition, p. 243 (Ex. S-8). Questions written by minority group persons have been incorporated into forms of the Commons over the past five years. Manning deposition, written questions on re-direct examination, Q. 40-41 (Ex. S-8).

The Committees of Examiners that create specifications and review questions for the Area Examinations that are relevant to portions of the general education tests include minority group members, as do the panels of question writers who create and review all proposed questions for these Area Examinations. These experts are identified in Exhibits S-MAN-4 and S-MAN-7.

³⁴ Exhibit S-MAN-29, entitled "ETS Minority Staff Members," identifies minority persons who work on the NTE tests, and outlines their professional credentials. Manning deposition, pp. 242-43 (Ex. S-8).

³⁵ Exhibit S-MAN-30 identifies minority group persons who have participated in ETS internship programs and who have reviewed NTE tests in the course of their work at ETS. Manning deposition, pp. 243-44, written questions on re-direct, Q. 33-34 (Ex. S-8).

³⁶ Exhibit P-MAN-11, entitled "Some Highlights of Minority-Related Research at ETS and CEEB [College Entrance Examination Board]" (March, 1975), contains brief descriptions of these studies. Manning deposition, p. 246 (Ex. S-8).

ciated with test-taking ability or experience is limited because an ample number of sample problems are provided in the NTE program publications that are mailed to candidates prior to the examination and clear unambiguous instructions are provided to candidates with regard to strategy involving speed and accuracy in taking the test (*e.g.*, skipping and omitting items), how the test is scored with respect to guessing, and the procedures for changing answers (*e.g.*, erasures).³⁷

The specifications of the Common Examinations were reviewed again in 1973 after ETS completed another broad survey of the curriculum requirements of teacher education institutions in cooperation with the American Association of Colleges for Teacher Education.³⁸ This survey covered 719 teacher training institutions, nearly 60% of all of such institutions in the United States. Data were collected on required subjects and course hours in general education for both elementary and secondary teachers. Similar data were collected on professional education course requirements, subdivided by psychological and sociological foundations of education. This study concluded: (1) that there had been relatively little shift in course emphasis since the 1963 revision of the Commons; (2) that the coverage of the tests in the Common Examinations is closely related to the current curriculum requirements of teacher training institutions; and (3) that the relative emphasis (as defined by the weighting of the four tests) fairly reflects the current academic program.

Test A: Of the 719 institutions studied, 645 (90%) have at least one course requirement in psychological foundations in education for elementary teachers and 665 (92%) have such a requirement for secondary teachers. Similarly, in the area of sociological foundations, 595 (83%) have at least one requirement for elementary teachers and 609 (85%) have at least one requirement for secondary teachers. The specific courses currently required by most teacher training institutions in the area of psychological foundations—general psychology, child psychology, educational psychology, human growth and development, and tests and measurement—are the subjects covered by the specifications for that portion

³⁷ Manning deposition, p. 235 (Ex. S-8). See also sample forms of the tests. Ex. S-MAN-27, S-MAN-28, S-MAN-28A.

³⁸ S. Sherwin, *TEACHER EDUCATION: A STATUS REPORT* (1974) (Ex. P-MAN-12) (hereinafter "Sherwin study").

of Test A. Similarly, in the area of social foundations, the required courses—introduction to education, philosophy of education, history of education, school in American society, and foundations of education—match closely the specifications for the corresponding portion of Test A.³⁹

Test B: The subject matter of Test B, written English expression, is required by an overwhelming 87% of the institutions surveyed.⁴⁰ In some areas, the percentage is even higher: *e.g.*, South Atlantic (93%), East South Central (94%), and West South Central (98%).⁴¹

Test C: The subjects covered by Test C are more widely varied than any other test, but data from the survey indicate a widespread similarity in requirements. For example, the requirements for elementary education majors generally included art (66%), geography (44%), government (44%), history (75%), and music (63%). Interestingly, and contrary to a fundamental proposition relied on (but not proved) by plaintiffs, black institutions consistently show a higher percent requiring specific subjects. For example, although only 66% of all institutions in the study sample require a course in art, 81% of the black institutions in that sample have such a requirement.⁴²

Test D: There was a similarly high degree of agreement on the utility of a general background in science and mathematics. For example, the required courses for elementary education majors included biological science (70%), physical science (67%), and mathematics (82%). In some areas there was virtual unanimity on the importance of a general background in mathematics (*e.g.*, South Atlantic (90%), East South Central (100%), and West South Central 94%). Again, black institutions had a significantly higher percentage requiring specific subjects in these areas. For example, biological science was a required course at 70% of all institutions

³⁹ Sherwin study, p. 19 (Ex. P-MAN-12).

⁴⁰ *Id.* at 11. Plaintiffs' expert Dr. Cohen also agrees that a test of written English expression is "of great importance in some sort of quality control of teachers." Cohen deposition, pp. 22-23 (Ex. P-111).

⁴¹ Sherwin study, p. 12 (Ex. P-MAN-12).

⁴² An institution was considered black if the number of blacks enrolled in the teacher training program was 70% or more of the total number of students enrolled in teacher education. There were 31 such institutions included in the survey. *Id.* at 6.

and at 97% of black institutions. Similarly, mathematics was required at 82% of all institutions and 100% of black institutions.⁴³

ETS has a very extensive program of research in teacher education and evaluation. These major research projects include investigations into teacher support programs (providing novice teachers with help and evaluating the utility of that help with respect to teacher effectiveness); teacher-student interaction; teacher evaluation and measures of teacher behavior; teacher attitudes; teacher effectiveness and surveys of the input and output in teacher education.⁴⁴ The results of this research are available to and used by the program and test development staff that work with the NTE and ensure that the NTE are up-to-date with respect to developments in the field.

4. *Computation of Scores*

The answer sheets completed by candidates are sent directly from the test center to ETS for processing. Before any scores are computed, ETS conducts statistical analyses of each question to determine whether the expected proportion of the total number of candidates who took the test selected the correct answer. Any question that produces an unexpected result—that is, many fewer candidates indicate the correct answer than should be the case, given the expected difficulty of the question—is re-examined by a test development specialist and unless a special justification is available, the question is then dropped from the examination and is not scored for any candidate.⁴⁵

ETS computes four scores for each of the four tests in the Commons: a raw score; a formula score (to correct for guessing); a scale score (to permit the score to be compared with other scores nationwide and over time); and a weighted score (that determines the contribution of each of the four tests to the final composite score).

The raw score is a simple count of the candidate's performance: the number right, the number wrong and the number omitted.⁴⁶

⁴³ Sherwin study, pp. 11-12 (Ex. P-MAN-12).

⁴⁴ A detailed description of twelve recent major research projects in these areas is set out in the record. Manning deposition, written questions on re-direct examination, Q. 44 (Ex. S-8).

⁴⁵ Manning deposition, pp. 82-83 (Ex. S-8).

⁴⁶ *Id.* at 151.

The formula score is computed by subtracting from the total number of right answers, one-fourth of the number of wrong answers. These tests all have a five-choice answer format. If a candidate guessed at all the questions and had no knowledge with which to eliminate one or more of the wrong answers, it would be expected statistically that one-fifth of the guesses would be correct and four-fifths would be incorrect. The candidate's formula score would be zero. This correction for guessing results in a more realistic estimate of the test-taker's true score.⁴⁷

The scaled score equates the level of difficulty of the current test with previous tests and converts the formula score to the standard scale used by ETS for this examination which runs from 250 to 990 points.⁴⁸ This conversion is done according to standard mathematical formulae. Scaling permits comparison of a candidate's score from a test administered in 1975 with the same (or a different) candidate's score from a test administered in 1965 or any other year.⁴⁹

The weighted score reflects the relative emphasis that the various subject areas are given within teacher training programs throughout the country. The professional education test is weighted 4.0 and accounts for 40% of the total score; Test B (written

⁴⁷ Manning deposition, pp. 151, 428-30 (Ex. S-8). The problem of how to deal with guessing is discussed in detail in *EDUCATIONAL MEASUREMENT*, pp. 59-61 (R. Thorndike ed. 1971).

⁴⁸ The score scale for the Common Examinations was originally set in 1940. At that time the mean of Commons scores for the scaling group was set at 600 and the standard deviation was set at 100. On a scale so defined, nearly all scores actually obtained by the scaling group would be expected to fall between 300 and 900. Until recently, scaled scores below 300 were reported as 300 and scaled scores above 900 were reported as 900. At the present time, scaled scores between 250 and 290, and between 910 and 990 are reported as such. Scores below 250 are reported as 250 and scores above 990 are reported as 990.

The reason for choosing a score scale such as that used on the NTE is that it is desirable to convert raw score data to scales with preassigned characteristics in round numbers that are easy to recall and easy to use. The choice of one range of reportable numbers rather than another is essentially arbitrary. ETS considers it desirable to avoid score scales that might easily be misinterpreted as percentages. The NTE score scale encourages a test user to rely upon published literature concerning the test in interpreting scores. Manning deposition, written questions on cross examination (Set I), Q. 1-6 (Ex. S-8).

⁴⁹ Manning deposition, pp. 152, 430 (Ex. S-8). See also Angoff, *Scales, Norms and Equivalent Scores*, in *EDUCATIONAL MEASUREMENT*, pp. 562-68 (R. Thorndike ed. 1971).

English expression) is weighted 1.0 and accounts for 10% of the total score; Test C (social studies, literature and fine arts) and Test D (science and mathematics) are each weighted 2.5, and each account for 25% of the total score.⁵⁰

5. *Properties of the Tests*

There are a number of measurement concepts by which certain qualities or characteristics of a test can be judged. The most important of these are standard error of measurement, reliability, speededness and validity. Under each of these benchmarks, the Common Examinations are fair, soundly constructed, and appropriate for the purposes for which they were designed.

a) *Standard error of measurement*

The measurement provided by a single score on a standardized test is to some degree imprecise as a measurement of the level of achievement the test taker has actually reached. The standard error of measurement describes the range within which a test taker's actual score would likely fall.⁵¹ The smaller the standard error, the more precise the measurement afforded by the test. The standard error for each of the tests that make up the Common Examinations is quite small: Test A, 3.4 points; Test B, 4.0 points; Test C, 3.4 points; and Test D, 3.9 points. The standard error for the Commons as a whole is 19 points.⁵²

This standard error remains relatively constant for these tests regardless of the heterogeneity or homogeneity of the group that

⁵⁰ Manning deposition, pp. 152-53, 432-34, 436-37 (Ex. S-8). The propriety of the weights was reaffirmed recently by the Sherwin study (Ex. P-MAN-12) which showed that most institutions continue to devote more hours to general education than to professional education. About 68% of the institutions surveyed required from 31 to 60 hours of general education, while 76% require from 16 to 25 hours of professional education. *Id.* at 14-15.

⁵¹ If an individual were to take equivalent forms of the test over and over again, his true score would be the average of all the actual scores he obtained. The actual scores would be dispersed around that average score and the measure of that dispersion around the average would be the standard error of measurement. Manning deposition, p. 157 (Ex. S-8). The formula for calculating the standard error is set out at Manning deposition, written questions on cross-examination (Set II), Q. 2 (Ex. S-8).

⁵² NTE TECHNICAL HANDBOOK, Table 4, p. 30 (Ex. S-MAN-9); Manning deposition, pp. 480-81, 483-84 (Ex. S-8).

takes the test. It would be virtually the same for members of minority groups and members of majority groups.⁵³

b) *Reliability*

Reliability is a statistical concept that is used to indicate consistency or stability of test results over time. It is usually expressed as a coefficient that ranges from zero to 1.0. The closer the reliability coefficient is to 1.0, the closer the test comes to perfect reliability which would occur if there were a perfect correlation between the scores obtained on one administration of the test and other administrations of equivalent forms of the test. There are a number of statistical methods for calculating the reliability of a test. ETS uses a method known as the Kuder-Richardson formula which is found in most standard texts.⁵⁴ For formula-scored tests such as the NTE, ETS uses the Dressel modification of the Kuder-Richardson formula 20.⁵⁵ ETS has compared results obtained by the Dressel modification of the Kuder-Richardson formula 20 with the results of three other methods of estimating reliability (alternate form correlations and odd-even correlations adjusted by the Spearman-Brown formula, both based on parallel forms administered at the same sitting, and alternate form correlations for tests administered six to ten months apart.)⁵⁶ All the comparisons justify the use of the Kuder-Richardson procedure with tests that show no more than moderate speededness.

The reliabilities of the tests that are part of the Common Examinations are quite high: Test A, .91; Test B, .88; Test C, .89; and Test D, .87. The reliability of the Common Examinations as a whole is .96.⁵⁷

⁵³ Manning deposition, p. 166, 484-88 (Ex. S-8).

⁵⁴ See, e.g., H. Gulliksen, *THEORY OF MENTAL TESTS* (1950); F. Lord & M. Novick, *STATISTICAL THEORIES OF MENTAL TEST SCORES* (1968); Stanley, *Reliability*, in *EDUCATIONAL MEASUREMENT* (R. Thorndike ed. 1971).

⁵⁵ Dressel, *Some Remarks on the Kuder-Richardson Reliability Coefficient*, 5 *PSYCHOMETRIKA* 305-10 (1940).

⁵⁶ Swineford, *An Assessment of the Kuder-Richardson Formula (20) Reliability Estimate for Moderately Speeded Tests*, paper presented at a meeting of the National Council on Measurement in Education (February, 1973) (Ex. S-MAN-12).

⁵⁷ List of Reliability Coefficients for National Teacher Examinations (Ex. S-MAN-11); Manning deposition, p. 176 (Ex. S-8).

These coefficients of reliability are not sensitive to the cultural and ethnic background of the group that took the tests, and generally psychometricians do not attempt to determine the reliability of a test by race or ethnic group.⁵⁸

c) *Speededness*

Speededness is a statistical concept that expresses the degree to which time is a factor in the performance of a sample of candidates on the test. If a candidate does not have enough time to finish the test, the test score may be a more imperfect measure of the candidate's actual level of achievement. A test is regarded as unspeeded if 80 percent of the candidates taking the test reach the last item, and if virtually all the candidates complete 75 percent of the questions on the test.⁵⁹

By that standard, Tests A, B and C of the Commons are unspeeded; and Test D is only slightly speeded.⁶⁰

d) *Validity*

The validity of a test is generally defined as the extent to which a test measures what it purports to measure. The basic underlying validity of the Common Examinations is established if the tests actually measure the content of the academic preparation of prospective teachers with respect to professional education and general education subjects. This is a form of "content validity," and it is established by the method by which the tests are constructed.

The American Psychological Association has described this professionally-accepted process for establishing the validity of a test of knowledge:

To demonstrate the content validity of a set of test scores, one must show that the behaviors demonstrated in testing constitute a representative sample of behaviors to be exhibited in a desired performance domain. Definitions of the performance

⁵⁸ Manning deposition, p. 170 (Ex. S-8).

⁵⁹ *Id.* at 172.

⁶⁰ NTE TECHNICAL HANDBOOK 27, 29 (Ex. S-MAN-9); Manning deposition, pp. 172-73, 286-87 (Ex. S-8). Candidates are informed in the instructions for the examination that there are time limits and those limits are printed at the beginning of every test. Manning deposition, p. 174 (Ex. S-8); *see, e.g.*, sample tests marked as Exhibits S-MAN-27, S-MAN-28, and S-MAN-28A.

domain, the user's objectives, and the method of sampling are critical to claims of content validity. An investigation of content validity requires that the test developer or test user specify his objectives and carefully define the performance domain in light of those objectives. APA STANDARDS at 28.

By the process described in Section 2 above, ETS has followed precisely the procedures described in the APA STANDARDS. The performance domain of each test has been carefully specified after detailed examination of the curricula of a large number of teacher training institutions and consultations with experts in elementary and secondary teaching as well as teacher preparation. Each test has a discrete objective that corresponds to a particular part of the curriculum.

Further, as described in Section 2 above, the method of sampling is carefully and professionally controlled, as required by the APA STANDARDS. A matrix is created to define the subject matter specifications of the test within each performance domain. The matrix is supplemented by additional specifications that control the number of questions in each subject and the level of difficulty of the questions.⁶¹

This method employed by ETS to construct the Common Examinations is described with approval in the APA STANDARDS:

If a test is used to estimate achievement in American history in Grade 12, the performance domain is less objectively defined [than in a previous example.] Given agreement on instructional objectives, it could be defined in terms of the types and quantities of the skills, facts and concepts of American history, as determined by the pooled judgments of authorities, experienced teachers, and competent curriculum makers in that field. APA STANDARDS at 28.

It is precisely such a pooled judgment of experts that is used to construct each part of the Common Examinations.

In addition, ETS has done two large studies to determine whether NTE scores are correlated with undergraduate grade point average, another measure of subject matter knowledge. In 1962, a study by Pitcher collected data from 11 colleges and universities in order to compare grade point averages with NTE scores. The aver-

⁶¹ Manning deposition, pp. 213-15 (Ex. S-8).

age correlation between overall grades and weighted Commons scores was .57. The average correlation between grades in professional education courses and scores on Test A of the Commons (professional education) was .52.⁶²

In 1975, a study by Wexler⁶³ considered data from over 30,000 candidates who took the NTE in November 1973 and January 1974. This was a concurrent validity study that examined the relationship between grade point averages and NTE scores for groups of students who took 24 of the 27 area examinations for which there was a sufficient sample size. The study results indicate that the probable lower bound of the correlation between grade point average and scores on the Common Examinations is about .36. The comparable correlation between grade point average and scores on the Area Examinations is about .34.⁶⁴

In addition, a research study done as a doctoral dissertation in 1969 gathered data on 249 students who had completed a secondary teacher education curriculum at a predominantly black college in the South. That study concluded

⁶² B. Pitcher, *The Relation of Academic Success in Teacher Preparatory Curricula to Scores on the NTE Common Examinations* (1962) (Ex. S-MAN-21). The professional credentials of the author are set out in Exhibit S-MAN-22. Dr. Manning testified:

"I believe [the study] is a reliable source that has been written primarily for professionals in the field of measurement. The publication has been, as is our practice, circulated to other professional staff within ETS and within the professional community. It has been subject to scrutiny for inaccuracy This is a good report authored by a highly competent person." Manning deposition, p. 228 (Ex. S-8).

No other witness testified with respect to the Pitcher study.

⁶³ N. Wexler, *Concurrent Validity of the National Teacher Examinations* (1975) (Ex. S-MAN-19). The professional credentials of Dr. Wexler are set out in Exhibit S-MAN-20. Dr. Manning testified:

"I believe that [the Wexler study] is a reliable source. It has been written primarily for a professional audience of persons in the test and measurements field. Dr. Wexler presented this paper at the meeting of the National Council on Measurement in Education in March of 1975 I believe it is a good piece of work and that he has done a highly professional job in authoring this report." Manning deposition, p. 226 (Ex. S-8).

No other witness testified with respect to the Wexler study.

⁶⁴ Wexler study, pp. 2-5 (Ex. S-MAN-19).

when a general test of knowledge of principles of teaching is needed, it seems that the National Teacher Examinations is a valid test.⁶⁵

6. Use

The Common Examinations are widely used. In 1975, 67,652 candidates took the Commons in all 50 states. Over 1,000 candidates took the Commons in each of 18 states.⁶⁶ This reflects, to some extent, state and local requirements with respect to the NTE. Georgia and West Virginia collect NTE data for general advisory purposes. Mississippi, South Carolina, Connecticut, Delaware, Hawaii, Montana, Oregon, Rhode Island and Tennessee use NTE data in certain state-wide certification functions. Several large school systems also use NTE data for certain certification or selection purposes: Philadelphia, Pa.; New York City, N.Y.; Boston, Mass.; Los Angeles, Calif.; Chicago, Ill.⁶⁷

In addition, the NTE are widely used in evaluating approved programs. In 1975 scores were reported by ETS to over 800 educational institutions that use these scores to evaluate their instructional programs, admission policies, grading procedures and student eligibility policies.⁶⁸ Scores on the Commons can be compared across institutions, therefore, the average scores of all students in a particular college who took the tests at one or more administrations can be compared with the average scores for institutions across the nation. This may indicate strengths and weaknesses in a particular academic

⁶⁵ E. Davis, *Relationships Between the National Teacher Examinations, Certain Variables, and Secondary Teacher Education Curricula*, Ch. III (1969) (Ex. S-MAN-41).

⁶⁶ Alabama, Arkansas, California, Georgia, Illinois, Kentucky, Louisiana, Massachusetts, Mississippi, New Jersey, New York, North Carolina, Pennsylvania, South Carolina, Tennessee, Texas, Virginia, West Virginia. Manning deposition, written questions on re-direct examination, Q. 55; written questions on re-cross examination, Q. 17 (Ex. S-8).

⁶⁷ A number of these uses involve cut-off scores. For example, Mississippi has established a state-wide score requirement of 850. Chicago has a cut-off score of 1100 for all public school teachers and Los Angeles uses a score of 1100 as an alternative means of certifying elementary school teachers. Manning deposition, written questions on re-direct examination, Q. 90, 91; written questions on re-cross examination, Q. 31, 32 (Ex. S-8).

⁶⁸ Manning deposition, written questions on re-direct examination, Q. 92 (Ex. S-8). ETS sends scores to institutions only at the request of candidates, therefore this large number of institutions receiving score reports reflects policies that require the submission of scores by students.

program or may help college officials to assess the level of the academic program that is most useful for a particular type of student body.

B. Rebuttal of Plaintiffs' Contentions with Respect to Disqualification of the Common Examinations

Plaintiffs' experts assert that the Commons should be disqualified as a basis for certification decisions because of basic flaws in the tests, regardless of the outcome of any validity study. They advance several contentions with respect to the Common Examinations, most of which were briefed and argued in the *North Carolina* case. It makes sense analytically to consider the underlying characteristics of the Commons apart from any particular use and the evidence establishing the validity of that use. When separated from the primary issues in this case with respect to the validity study, the plaintiffs' arguments relating to the underlying properties or characteristics of the Common Examinations can be summarized in a logical series as follows: The tests are racially or culturally biased; if not biased, the tests are centered on the wrong objective; if properly directed, the tests are out-of-date; and if not out-of-date, the tests are not fairly representative of teacher training curricula.

When stated in a series of steps of diminishing importance as set out above, and classified logically instead of as make-weight support for unrelated arguments with respect to the validity study, these assertions and related evidence are readily met.

1. *The Tests Are Not Racially or Culturally Biased*

Plaintiffs rely on statistical evidence with respect to scores achieved by blacks and whites, two studies, and the testimony of one expert witness to suggest the possibility of racial bias. None of this evidence supports that contention.

(a) *Statistical compilations of scores*

The statistical evidence with respect to the scores achieved by blacks and whites on the Common Examinations indicates that blacks generally do more poorly than whites. That evidence is presented by

plaintiffs in many different analyses of relative performance and undoubtedly reflects some measure of reality.

But the analysis of the impact of race on test scores has never been that simple. The APA STANDARDS state flatly:

A simple difference in group means does *not* by itself identify an unfair test (p. 43) (Emphasis in original.)

Similarly, the DIVERSITY 14 PRINCIPLES assert:

A simple difference between groups, whether in selection ratios, mean scores, or correlation coefficients, is *not* adequate evidence of bias in the use of an assessment procedure (p. 2) (Emphasis in original.)⁶⁹

There is a strong inference to be drawn from the statistics that the education of blacks is less thorough or effective than the education of whites in South Carolina. A difference in education patterns is readily apparent from the record. There are 25 teacher training institutions in South Carolina and each has a predominantly black or predominantly white student body. Further, these teacher training institutions have different admissions policies and draw students from many different preparatory schools. Some schools with very high standards take only the most qualified applicants and need do virtually no remedial work, so all of the teacher training program can be devoted to advanced learning. Other schools have open admissions policies and must devote considerable time to remedial work, leaving relatively less time for advanced subjects. Some schools have greater resources and are able to offer a diverse curriculum through a highly qualified faculty. Other institutions must make do with less. The differential effect on education of relatively greater or lesser financial resources has been amply documented before the courts.

Further, there is a less direct inference to be drawn about the effect of socioeconomic status on the ability to absorb academic training. Many students at institutions with open admissions policies and very low tuition come from families with lower socioeconomic status and are restricted in their choice of educational institution by financial considerations. The barrier to effective education created by poverty is well-documented in educational research.

⁶⁹ See also Manning deposition, pp. 296-97, 301-03 (Ex. S-8).

The Common Examinations are all knowledge oriented tests—achievement tests rather than aptitude tests. If an examinee does not have the knowledge called for by the question, he cannot answer the question no matter how much ability or aptitude he may have. Guessing does not help his score because ETS uses a formula score that corrects for guessing. If the educational institution that the examinee attends does not impart the knowledge, or enough of the knowledge required to answer correctly the questions on the Commons, then a low score will result regardless of the race of the examinee.⁷⁰

None of the data collected by plaintiffs show that blacks with high school and college training that is equal, in all substantive respects, to the high school and college training of whites, still receive lower scores on the Commons. None of the data collected by plaintiffs show that blacks with socioeconomic status that is equal, in all substantive respects, to the socioeconomic status of whites, still receive lower scores on the Commons.⁷¹ None of the data submitted for the record in this case, standing alone, supports only the inference that the Common Examinations are somehow inherently biased against minority group persons.

(b) *The Spencer study*

In 1972, Thelma Spencer completed a doctoral dissertation entitled "An Investigation of the National Teacher Examinations for Bias with Respect to Black Candidates."⁷² This dissertation was completed while Dr. Spencer was an ETS employee, but was prepared for submission to the faculty of the graduate school of the University of Colorado and was not an ETS research project.

Dr. Spencer analyzed score data from the January 1971 administration of the Common Examinations for samples of candidates drawn from Northern and Southern teacher training institutions and from public and private examples of each. Predictably, the data

⁷⁰ Manning deposition, pp. 249-52 (Ex. S-8).

⁷¹ Data showing that blacks attending teacher training institutions that have a predominantly white student body achieve lower average scores than whites attending that same institution do not make this showing because they fail to take into account the differential preparation for college that may have been experienced by blacks and whites at the same institution or the different socioeconomic status of the candidates. These factors make a great deal of difference with respect to the amount of the teacher training curriculum that can be mastered in the four-year time limitation.

⁷² This document is Exhibit P-MAN-21 (hereinafter "Spencer study").

showed that white candidates from public and private institutions in the South performed better than black candidates from public and private institutions in the South. The data also showed that candidates from integrated public and private teacher training institutions in the North did better on the Commons than did black candidates from public and private institutions in the South.⁷³

The study pinpointed some of the reasons for this score differential. It was found that Test B, in written English expression, presented less than average difficulty for white Southern candidates and candidates from integrated Northern institutions. However this test was highly difficult for Southern black candidates because of their poor performance on the English usage questions.⁷⁴ The study showed that Test D, in science and mathematics, was "quite difficult for black candidates, of average difficulty for the white candidates, and below average difficulty for the Northern candidates."⁷⁵

The findings of the Spencer study with respect to differential performance by race do not support a conclusion that the tests are racially biased. Differential performance by itself is not a professionally accepted definition of bias.⁷⁶ The key questions with respect to the racial bias of a test are: (1) whether the knowledge necessary to answer the questions correctly is fairly related to the objective of the test; and (2) if so, whether, in a comparison of blacks and whites who both *know* the material required to answer the question, blacks systematically do more poorly than whites. If the critical dimension of *knowledge* is left out of the analysis, the statistics are meaningless for purposes of determining the existence of bias.⁷⁷ None

⁷³ Spencer study, pp. 228-29 (Ex. P-MAN-21).

⁷⁴ *Id.* at 93.

Some critics argue that blacks do more poorly on English usage questions because blacks typically do not speak "standard" English, therefore the test is "biased" against them. However, it seems reasonable to require persons who aspire to teach to demonstrate mastery of "standard" English regardless of their personal preference with respect to dialect.

⁷⁵ *Id.* at 95.

Plaintiffs' experts concede that it is reasonable to require teachers to be able to do basic arithmetic and understand elementary aspects of science. Cohen deposition, p. 66 (Ex. P-111).

⁷⁶ See discussion at pp. A-23 to A-25, *infra*.

⁷⁷ The relatively lower scores of blacks on the mathematics test confirm the need to use this approach. There are no "black-oriented" or "white culture" questions on the mathematics test, just problems in mathematics. Manning deposition, pp. 240-41 (discussing Exhibit S-MAN-28A which is a sample of the Area

of the statistical comparisons offered by Dr. Spencer addresses either of these questions.

The Spencer study included a second and distinct effort to analyze racial bias. Dr. Spencer assembled a panel of five judges—a high school principal, two classroom teachers, a guidance teacher and a clinical psychologist—to review the questions on the Commons for possible racial bias.⁷⁸ The definition of bias constructed and used by the panel was as follows:

Biased Test Item. An item which meets one or more of the following criteria: (1) includes terms which are offensive, patronizing or demeaning to blacks; (2) includes experiences not generally available to blacks; (3) includes pejorative euphemisms and stereotypes.⁷⁹

The five judges reviewed each of the 270 questions on the Common Examinations and found 14 that met this very broad definition. The precise findings are instructive of the sweeping application of the concept of bias:

In reviewing the content of items rated as biased by the judges, it was possible to classify the 14 items to some extent, and to determine which criterion fit a particular item.

Item A₁ was possibly perceived as including pejorative euphemisms and stereotypes. The same criterion could be applied to items A₂, A₃, B₁, B₂, C₁. This group of items contains phrases like "lower-class children," "a ghetto child," "disadvantaged homes," and "black power."

The criterion of terms which are offensive, patronizing or demeaning to blacks was met in items A₄ and A₅. Phrases which have taken on the characteristics of code words are included in these two items: "busing to achieve racial integration" and "racial inequality."

The last seven items, C₂, C₃,¹ C₄, C₅, C₆ and D₁, fit the criterion of including experiences not generally available to blacks. The items in test C ask for knowledge about art and literature which are generally believed to be a reflection of the majority culture

Examination in Mathematics). The compelling inference to be drawn from the differential performance by race in mathematics is a difference in knowledge of mathematics, probably stemming from a difference in educational opportunity with respect to mathematics.

⁷⁸ Spencer study, p. 84 (Ex. P-MAN-21).

⁷⁹ *Id.* at 15.

and, as a consequence, not of interest to or a part of the experience of many black people. This is more true of the items in test C than of item D₁ which is a science item. The latter was included, perhaps, because it was perceived as a very difficult item which asked for knowledge which is relatively unknown to the average teacher candidate.⁸⁰

These findings of the panel of five judges selected by Dr. Spencer do not support a conclusion that the tests are biased for a number of reasons:

(1) These findings are contrary to the findings of other reviewers who were minority group persons and who examined the tests for possible bias. For example, in 1969, two years prior to the Spencer study, a minority review panel of seven members considered the phrasing of each of the questions on the Commons and did not find terms such as "busing to achieve racial integration," "racial inequality," "disadvantaged homes" or "black power" to be offensive, patronizing, demeaning or stereotypical.⁸¹

(2) As the Spencer study concludes, the parts of the definition of bias having to do with the use of particular words do not affect score levels. The study points out:

[I]t was noted that items which were perceived as offensive to blacks or including pejorative euphemisms and stereotypes *were no more difficult for black candidates* in the final analysis. In other words, even though an item might be perceived as biased, it does not necessarily follow that blacks will get it wrong and whites will get it right. It appears that candidates respond more readily to such items on the basis of their knowledge of item content than on the basis of race.⁸² (Emphasis added.)

(3) The definition of bias includes questions that are somehow directed at "experiences not generally available to blacks." That definition is not useful for analyzing the NTE because it is missing two critical elements: (a) the relationship of the question to the performance domain to be tested; and (b) the reasonableness of the requirement with respect to that performance domain. A test in

⁸⁰ Spencer study, p. 220 (Ex. P-MAN-21).

⁸¹ ETS removed each of these questions from subsequent versions of the test, despite its view that the criticism was over-extended. Manning deposition, p. 248 (Ex. S-8).

⁸² Spencer study, pp. 224-25 (Ex. P-MAN-21).

Spanish is not biased just because training in Spanish is not an experience generally available to blacks. It is reasonable to require candidates who seek certification as Spanish teachers to demonstrate proficiency in Spanish whether or not that is a "experience" generally available to persons of the ethnic group to which the candidates belong.

(4) The assumption that questions in Test C, particularly with respect to literature and fine arts, are "a reflection of the majority culture and, as a consequence, not of interest to or a part of the experience of many black people" is unsupported by any evidence and probably unwarranted. The Sherwin study revealed that black institutions generally have more extensive requirements in these areas than do white institutions and black students are therefore more likely to be exposed to these materials in the course of their academic training than are whites. Black institutions typically include more general education in the curriculum and 52% of these institutions favor the junior year as the point of entry into teacher education. The corresponding percentage for all institutions was 31%.⁸³ Further, the black institutions consistently show a higher percent requiring specific subjects in general education. For example, only 66% of all institutions surveyed had a course requirement in art, but 81% of the black institutions had such a requirement.⁸⁴

(5) The inclusion of a science question in the domain of "experiences not generally available to blacks" is particularly arbitrary and Dr. Spencer's rationale that this was a difficult item for *all* teacher candidates is plainly irrelevant to a determination of bias.⁸⁵

(c) *The Medley study*

The Medley study was begun a year before the Spencer study, but covered administrations of the Common Examinations in two successive years and thus was to some extent concurrent with the Spencer study.⁸⁶ The Medley study had a narrower focus than

⁸³ Sherwin study, p. 10 (Ex. P-MAN-12).

⁸⁴ *Id.* at 11-12.

⁸⁵ The Sherwin study indicates that black teacher training institutions typically have more requirements with respect to science courses. For example, 70% of all institutions surveyed required study of biological sciences, but 97% of the black institutions had such a requirement. *Id.* at 12.

⁸⁶ The Medley study was completed in 1971; first published in 1972, Medley *et al.*, *Race and Subject-Matter Influences on Performance on General Education*

the Spencer study; it covered only Test C, the social studies, literature and fine arts test, which contains 65 questions out of a total of 270 questions on the entire Commons. This study examined the extent to which score levels of black candidates on Test C would increase if "black" and "modern" questions were substituted for the then-current questions.⁸⁷

The study finds that substitution of "black" questions for "traditional" questions would result in a maximum increase in the average score of the average black candidate relative to that of the average white candidate of about 19 points (out of a total of 60) or nearly 30%.⁸⁸ A change from a test composed of all "traditional" questions to a test composed of half "black" and half "modern" questions would increase the average score of black candidates by 12 points or about 20%.⁸⁹

These findings do not support a conclusion that Test C of the Commons is racially biased for several reasons:

(1) The finding that blacks are generally more familiar with "black culture" materials than with "majority culture" materials,

Items of the National Teacher Examinations, ETS Research Bulletin 72-43 (Sept. 1972) (Exhibit P-MED-14); and thereafter published in a shorter version in 1974. Medley, et al., *The Application of a Factorial Design to the Study of Cultural Bias in General Culture Items on the National Teacher Examination*, 11 JOURNAL OF EDUCATIONAL MEASUREMENT 235 (1974) (Ex. P-MED-12).

⁸⁷ The study offers the following definitions:

Items classified as black items dealt with aspects of the experience and culture contributed by black authors, artists, writers, musicians, and important historical and political black figures. Items classified as modern items dealt with aspects of contemporary culture and contemporary affairs such as recent developments in music, art, literature, social philosophy, social change, and politics, not oriented toward the black culture. Items classified as traditional items dealt with such topics as world trade, geographical aspects of countries, chronological development of various historical events, differences between political systems, artifacts of various cultures, authors and aspects of world literature, music and dance which did not emphasize aspects of either black or modern culture. Medley study, p. 236 (Ex. P-MED-12).

⁸⁸ There are a number of groups used in the study for comparison purposes. The increases in scores of black candidates ranged from 17 to 19 points for a shift from all "traditional" questions to all "black" questions; and from 11 to 13 points for a shift from all "traditional" questions to half "black" and half "modern" questions. Medley study, pp. 15, 33 (Ex. P-MED-14).

⁸⁹ *Ibid.*

does not support a conclusion that Test C is racially biased.⁹⁰ Presumably Polish candidates would do better on "Polish culture" questions and Jews would do better on "Jewish culture" items. The critical question is whether some general knowledge of "American culture" materials and their antecedents in other cultures is a fair requirement for teachers in public schools, regardless of whether the teachers themselves come from a so-called mainstream "American culture" environment. American culture undeniably includes contributions by blacks and other minority groups. Test C recognizes that contribution; as pointed out above, 15% of the questions on Test C call for knowledge with respect to minority culture or concerns.

(2) The contribution of Test C to the entire Commons score is not unfairly large—65 questions of a total of 270 (approximately 24%) and the Test C score is weighted as 25% of the total Commons score. Therefore the effect on the composite NTE score of the relative proportion of "black" questions on Test C is insignificant.⁹¹

(3) The change proposed by the Medley study, from "traditional" questions to "black" questions, changes the essential nature of the test. It is not a necessary conclusion that racial bias is present

⁹⁰ There is an alternative, and equally plausible, inference as to the differential results. This study was completed in 1971, several years after the initial drive to include "black studies" programs in college curricula. Colleges with predominantly black student bodies felt that pressure in particularly acute ways, and once such courses were made a part of the curriculum, there was a similar, and understandable, pressure on black students to take them. "Black studies" courses typically dealt with the contributions of black authors, artists, writers, musicians and important black historical and political figures. What the "black" questions in the Medley study were measuring was the differential propensity of black and white students to take "black studies" courses. Blacks did better because they had taken the course and obtained the specific knowledge called for by the questions. Manning deposition, written questions on re-direct examination (Set II), Q. 11 (Ex. S-8).

It is interesting to note that on the so-called "modern" items, black candidates performed more poorly than white candidates in all comparison groups, and that in three of the comparison groups, black candidates did even *more* poorly on "modern" items than they did on "traditional" items. Medley study, Table 5 (Ex. P-MED-12).

⁹¹ Dr. Medley conceded this point:

Q. Dr. Medley, what implications, if any, do the results of your study have for the results of a single cutoff score on the National Teacher Examinations on scores applied to both blacks and whites?"

A. Substantively, they don't have any great implication, I don't think, because it represents a relatively small part of the examination. Medley deposition, p. 74 (Ex. P-112).

in a test if changing the test to a different subject matter produces different results.

(4) The definition of "cultural bias" in the Medley study is significantly different than the definitions generally used by psychometricians and distinctly different than the use of the term in a legal context. Medley defined cultural bias simply by the content of the question.⁹² Under this definition, any question that does not deal specifically with "black culture" is biased against blacks. That definition may have utility for educational research, but it is not helpful in the legal context.

(d) *The Cohen opinion*

Dr. Elizabeth Cohen, a sociologist on the faculty at Stanford University, testified on behalf of plaintiffs and opined that one part of Test C, the fine arts section, is culturally biased. As a basis for that opinion, Dr. Cohen testified that she had inspected the Common Examinations on one occasion in 1974 and had found that the fine arts section contained certain items she defined as "too high culture, referring to a sort of European culture." In Dr. Cohen's opinion:

[B]lack people from a rural background would have very little access to this cultural information . . . and would therefore be put at a great disadvantage on the test because of the inclusion of such items. And I felt it was indefensible to have such items in the test. It constituted a cultural bias in the test to have that section, those items.⁹³

With respect to Dr. Cohen's opinion, it is important to note:

(1) Dr. Cohen is not an expert in any aspect of testing or measurement.⁹⁴ Therefore her opinion with respect to the cultural bias of the fine arts portion of Test C is entitled to no more weight than that of any layman.

(2) Dr. Cohen limits the possible adverse impact to blacks "from a rural background," indicating that socioeconomic factors may be more important than racial factors with respect to that opinion.

⁹² Medley deposition, p. 70 (Ex. P-112).

⁹³ Cohen deposition, p. 27 (Ex. P-111).

⁹⁴ Dr. Cohen so testified:

Q: Do you consider yourself an expert in the field of mental measurement?

A: I do not. *Id.* at 11.

(3) Dr. Cohen's observation is limited to the fine arts section of Test C. Only 15 of the 65 questions on Test C are devoted to fine arts,⁹⁵ and not all of these questions are devoted to "high culture" materials, even under Dr. Cohen's sweeping definition of anything that has to do with things European.⁹⁶

(4) There is evidence in the record that black colleges tend to give *more* emphasis to "high-culture" subjects, such as art, than white colleges.⁹⁷

(5) The minority review panel assembled by ETS in 1969 did not recommend deletion of all items that required knowledge of the European influences on American culture.

(6) Dr. Cohen's experience with the Common Examinations arises out of her appointment by ETS to the Teacher Programs Advisory Board, a group of 16 persons that advises ETS with respect to testing and research programs in the teaching area. Dr. Cohen's colleagues on that Board included six minority group persons⁹⁸ but plaintiffs choose not to depose any of them. The only fair inference is that the minority group members of the Board were of a different view than Dr. Cohen with respect to this issue.

2. The Tests Have a Legitimate Objective

Plaintiffs' experts contend that the only legitimate objective for tests that seek to measure any characteristic relevant to prospective teachers is on-the-job performance in the classroom.⁹⁹ Since the Common Examinations seek to measure performance in a training program rather than the ability of prospective teachers to teach effectively, these critics dismiss the Commons as focused on a wrong objective.

⁹⁵ Common Examinations, Test C (Form XNTI) pp. 27-33 (Ex. S-MAN-27).

⁹⁶ The black panel that participated in Dr. Spencer's study found only five such items using a similar test of bias (items involving experiences not generally available to blacks). Spencer study, p. 220 (Ex. P-MAN-21).

⁹⁷ See discussion of Sherwin study at pp. A-13 to A-15 *supra*.

⁹⁸ Exhibit S-MAN-5 identifies the members of the Teacher Programs Advisory Board and indicates which are minority group persons. Manning deposition, pp. 246, 292-93 (Ex. S-8). Plaintiffs did not choose Dr. Cohen as a witness for reasons of geographical convenience as she is a resident of California. Cohen deposition, p. 4 (Ex. P-111).

⁹⁹ *E.g.*, Cohen deposition, p. 31 (Ex. P-111).

This argument ignores two important and practical points. First, it is not possible, with present technology, to construct a test that measures the ability of prospective teachers to put to use in the classroom what they have learned in the training program, as a narrow objective, or to teach effectively, as a broader objective. The elements of effective teaching are complex and not easily measured by the type of test that can be administered to large numbers of candidates.¹⁰⁰

Second, although the knowledge gained in the teacher training program may not guarantee that a person will become an effective teacher, it is a reasonable proposition that a person is unlikely to be an effective teacher without such knowledge. Although the test cannot measure all the elements that contribute to effective teaching, it can measure one key element. The importance of academic preparation is not seriously questioned by the plaintiffs here. Every state has required, by statute or regulation, some amount of professional education in addition to preparation in the specific subject to be taught. A test that assesses achievement in this professional education program is like any other test that measures the outcome of a training program. As long as the training program is, on its face, rationally related to effective teaching, it is a reasonable objective for a standardized test to assess the outcome of that training program.

A variation of this criticism is that the input to the Commons comes from ETS staff experts and other teacher training program faculty members, not from classroom teachers. This, it is argued, focuses the test toward the elements necessary to succeed in the academic program rather than what is necessary for effective classroom teaching, which only teachers can really judge.

This argument has obvious flaws. First, it assumes that faculty members at teacher training institutions and subject matter experts have less of a grasp of what is necessary for effective teaching than classroom teachers. There is no support in the record for that proposition and common sense dictates otherwise. Classroom teachers tend to acquire a very narrow focus because of the nature of their classroom experience. They have "expert" qualifications only

¹⁰⁰ It may be possible to assess the ability to teach effectively by means of tests that are largely psychiatric interviews. This avenue has not been explored with any enthusiasm in educational research because of the prohibitive expense and evident impracticality of such tests for use with large numbers of candidates.

with respect to the needs of the particular grade in the particular location and setting in which they teach, and they have limited opportunities to observe the teaching performance of their colleagues. Faculty members of teacher training institutions, through their extended professional training, teaching experience and research, are more likely to have the kind of overview and perspective that is helpful in test construction.

Second, this argument assumes that teacher education experts have had no classroom teaching experience themselves. That is very often not the case. Many of the experts who serve on ETS committees or panels have been classroom teachers prior to or concurrently with their duties as faculty members at teacher training institutions.

Third, it assumes that the classroom teachers who do serve on ETS committees or panels are ineffective in expressing their point of view to the other members or are ignored by the other members, who constitute a majority. Plaintiffs have not examined any member of any committee of examiners or panel of question writers to attempt to demonstrate that assumption.

Further, an extensive research study conducted in 1975 by ETS in conjunction with the American Association of Colleges of Teacher Education demonstrates that classroom teachers have substantially the same view of the importance of the knowledge tested by the Commons as do those who constructed the tests.

This study surveyed 1500 classroom teachers in 10 states. The teachers were asked to rate 169 knowledge objectives in terms of their judgment of the importance of the knowledge for effective teaching. ETS staff members compared the ratings on knowledge objectives to the test content descriptions for the Common Examinations. This comparison identified 76 knowledge objectives rated essential by 51% or more of the teachers that were covered by the Commons, and 59 knowledge objectives rated moderately important by 51% or more of the teachers that were covered by the Commons. A further comparison indicated that 149 questions on the Common Examinations form XNT1 were related to knowledge objectives rated essential or very important by the teachers.¹⁰¹

¹⁰¹ B. Humphry *et al.*, ETS-AACTE PILOT STUDY: TEACHER RATINGS OF KNOWLEDGE OBJECTIVES OF UNDERGRADUATE TEACHER EDUCATION (1975) (Ex. P-EDE-10); Manning deposition, written questions on re-direct examination, Q. 39 (Ex. S-8).

3. *The Tests Are Up-To-Date*

Two of plaintiffs' experts, Drs. Cohen and Edelfeldt, criticized the Common Examinations as being out-of-date.¹⁰² Dr. Cohen, in particular, found Test A, the professional education test, to be deficient in this respect. She opined that (1) some items in the "social area reflected the predominant thinking of people in that area of the sixties . . ." and (2) "there was a tremendous emphasis on vocabulary of words that were currently popular in the '60's in teacher training courses."¹⁰³ Plaintiffs also point to references in ETS documents to plans for revising the Commons in various ways as support for the proposition that the Commons were (and are) out-of-date and in need of revision.

Plaintiffs' contentions in this regard should be rejected. Since 1963 when the Commons underwent a major revision, ETS has kept the examinations current. The specifications were reviewed in 1966 when the professional education subtests were combined, in 1969 when the minority review panel met, in 1973 when the Sherwin report on the current status of teacher education programs was completed, and again in 1976 after the ETS-AACTE survey of teachers with respect to important classroom knowledge was completed. New forms of the Common Examinations were constructed in *each* year from 1970 to 1975 utilizing special committees of outside question writers, experts in their fields, to keep the professional education test up-to-date, and drawing on the committees of examiners for the Area Examinations, also made up of outside subject matters experts, to keep up-to-date in the general education fields. Plaintiffs' experts have not specified a single question that is allegedly "out-of-date."

¹⁰² Drs. Cohen and Edelfeldt were repeatedly asked by plaintiffs to characterize their views as representative of the views of the other members on the Teacher Programs Advisory Board. *E.g.*, Cohen deposition, p. 15 (Ex. P-111). The Board is a group of 16 persons, broadly representative of differing viewpoints within the education community, brought together to advise ETS with respect to the NTE and other teacher-oriented testing and research programs. Such characterizations by Drs. Cohen and Edelfeldt are inadmissible hearsay, and were objected to on those grounds by the State. Further such characterizations are inaccurate. Drs. Cohen and Edelfeldt may have held such views, but they were not in a majority on the Board. Manning deposition, written questions on re-direct examination, Q. 128, 129 (Ex. S-8).

¹⁰³ Cohen deposition, pp. 27-28 (Ex. P-111).

Further, the opinions of plaintiffs' experts with respect to this matter are entitled to little weight. Dr. Cohen had no basis for offering an opinion that the Common Examinations are out-of-date except as compared to the curriculum at a single institution, Stanford University, which she characterizes as "very superior" with "course offerings . . . far enriched because in subject matter specialty [the students] study on the graduate level. . . ." ¹⁰⁴ The NTE are not designed to reflect what "very superior" teacher training institutions regard as "up-to-date." The examinations are broadly representative of the current offerings in teacher training curricula across the nation. Dr. Cohen had no basis for extending her criticism to that purpose, nor did she attempt to do so.

4. The Tests Are Fairly Representative

Plaintiffs have suggested, at various points in their construction of the evidentiary record, that the Commons, even if up-to-date and properly directed at assessing teacher training programs, are not fairly representative of those programs.

They point out that the Commons are used for assessment of the achievement of persons who are prospective elementary school teachers and also persons who are prospective high school teachers, and argue that the academic preparation for these two specialties is different and therefore the test must be unfair to one or the other.

Any assertion with respect to differential preparation of elementary and secondary school teachers as to the subjects included on the Common Examinations is effectively rebutted by the data assembled in the Sherwin study, completed in 1973. Comparisons of the mean number of course hours required for seven types of courses in psychological foundations and six types of courses in social foundations indicate no differences that would systematically affect performance on a test such as Test A. ¹⁰⁵ Indeed, there is remarkable similarity. This data is summarized in tabular form below.

¹⁰⁴ Cohen deposition, p. 49 (Ex. P-111).

¹⁰⁵ Sherwin study, pp. 18-19 (Ex. P-MAN-12).

Professional Education Required Courses

A. Psychological Foundations

	Elementary Education Mean Course Hours	Secondary Education Mean Course Hours
General Psychology	3.4	3.4
Child Psychology	3.1	2.7
Adolescent Psychology	2.7	3.0
Educational Psychology	3.2	3.2
Human Growth Dev.	3.4	3.4
Tests & Measurements	2.7	2.7
Other	3.6	3.6

B. Social Foundations

Introduction to Education	2.9	2.9
Philosophy of Education	2.9	2.9
History of Education	2.6	2.7
School in American Society	3.1	3.2
Foundations of Education	3.2	3.2
Other	3.2	3.2

Further, the number of course hours required in general education shows the same pattern in both elementary and secondary education fields:

Number of Hours	Elementary Teaching		Secondary Teaching	
	No. of Institutions	%	No. of Institutions	%
31-45	119	27%	109	25%
46-60	181	41%	189	43%
61-75	75	17%	53	12%

It is reasonable to conclude that the tests are proper for both elementary and secondary teacher candidates.

APPENDIX B

ANALYSIS OF PLAINTIFFS' CRITERION-RELATED VALIDITY
STUDY OF THE USE OF NTE SCORES BY SOUTH CAROLINA

Plaintiffs produced a validity study of their own, apparently intended to demonstrate that there was no rational relationship between NTE scores and on-the-job performance. A summary of the data collected in that study has been admitted into evidence by stipulation of the parties.¹

Although defendants stipulated to the admissibility of this data, the reasons that could have been advanced in opposition to admissibility are also relevant to the *weight* to be accorded this evidence. Therefore, it should be pointed out at the outset that both the supervisory ratings and the NTE scores that were used in the study are hearsay,² for which no foundation was laid. The record includes computer printouts reporting the results of statistical comparisons of these supervisory ratings and NTE scores. There is no indication that the data were collected correctly, that the method of summarization and statistical analysis was appropriate, or that the results are accurate tabulations. For these reasons, the court would normally give little or no weight to this evidence.

However, there is another purpose for which this court can consider this evidence—as a reasonable example of the feasibility of a criterion-related validity study under the circumstances in this case. In that light, plaintiffs' validity study demonstrates (1) that there is no usable criterion because available supervisory ratings, the best current possibility, are inherently unreliable; (2) even if the criterion were acceptable, there is an inadequate number of cases to permit a useful study; and (3) the steps that would have to be taken to convert the data into usable form lead necessarily to unreliable results.

(i) *Available supervisory ratings are inherently unreliable as a criterion.*

The supervisory ratings used by plaintiffs came from the records of three defendants alleged by plaintiffs to represent a defend-

¹ Stipulation No. 20. This stipulation covers the authenticity of plaintiffs' computer analysis.

² The availability of any exception to the hearsay rule with respect to any of these documents would depend upon individual circumstances as to which there is no information in the record.

ant class of school districts: Richland County, Charleston County, and Colleton County. Plaintiffs found that school districts within these counties use or have in their files 25 different kinds of evaluation forms; Richland has three, Charleston has 17, and Colleton has five.³ It is not known whether these 25 kinds of forms represent all of the forms used by these counties or are merely a sample picked by plaintiffs.

The "raw data" of plaintiffs' study confirms the opinion of most of the experts who testified in this case—that there is no generally accepted criterion of effective teaching. Each of the 25 kinds of forms reflects a different perception of what should be rated in order to comprehend fully the range of skills necessary to teach effectively. Further, the skills to be rated are defined quite differently. For example, some forms include a rating with respect to English usage which is variously phrased as "evidence of good cultural background (vocabulary, pronunciation)"⁴ or "the teacher uses appropriate language,"⁵ or "communication skills—ability to speak and write well,"⁶ or "voice and speech,"⁷ or "speech pattern,"⁸ or "demonstrates competency in English usage."⁹ These various formulations of a fairly narrowly defined quality with respect to teaching may or may not all mean the same thing. Further, using any one of those formulations, one rater may emphasize use of proper grammar, another may emphasize possession of an adequate vocabulary, and another may focus on the tendency to use slang, dialect or profanity.¹⁰ For these reasons, any individual teacher

³ Richland: Stipulation No. 20, Attachments B-1, B-3, B-6;

Charleston: *Id.* Attachments B-31, B-34, B-36, B-39, B-43, B-46, B-48, B-50, B-53, B-55, B-57, B-59, B-61, B-63, B-65, B-67, B-69;

Colleton: *Id.* Attachments B-12, B-15, B-18, B-20, B-23.

Plaintiffs apparently violated the confidentiality requirement with respect to these documents by leaving the name of the person rated on at least nine of the 27 sample forms submitted for the record. *Id.* B-2, B-4, B-6, B-15, B-23, B-28, B-46, B-51, B-65.

⁴ *Id.* Attachments B-1, B-3.

⁵ *Id.* Attachments B-12, B-16, B-55 ("Personal speech habits").

⁶ *Id.* B-28, B-6 ("Expresses self well").

⁷ *Id.* B-32, B-37.

⁸ *Id.* B-69.

⁹ *Id.* B-41, B-51 ("Displays proper use of English language and ability to express self orally"), B-57 ("Uses good English"), B-65 ("Use of English"), B-67 ("Skill in use of English").

¹⁰ Interestingly, at least 11 of the 25 rating forms did not consider language or English usage as an item important enough to be rated. *Id.* B-18, B-20, B-23,

B-3

might be rated differently by the same rater using the different formulations or differently by different raters using the same formulation. Plaintiffs' study provides no information about the raters or how they made their ratings. These variables are generally very difficult to control, and therefore most experts consider available supervisory ratings an inherently unreliable criterion.¹¹

(ii) *The number of cases is too small to permit a useful study.*

Even if one could define one or more criteria of effective teaching and provide an effective measure of such criteria, the number of cases available for study must be large enough. If that number is too small, the study may not be useful.¹²

Plaintiffs' study also illustrates this problem. A large majority of the hundreds of comparisons for which coefficients are given are of very small samples and the results are not statistically significant at the .05 level. For example, only nine out of 69 correlations between NTE scores and the various ratings of subject matter knowledge involve correlations that are statistically significant.¹³

B-26, B-34, B-43, B-46, B-53, B-59, B-61, B-63. Therefore, the rating forms (and presumably the raters) cannot even agree on the important items to be rated.

¹¹ *E.g.*, McNeil *et al.*, *The Assessment of Teacher Competence*, in *SECOND HANDBOOK OF RESEARCH ON TEACHING* 232 (Travers ed. 1973); D. Ryans, *CHARACTERISTICS OF TEACHERS* 43-44 (1960).

¹² *See, e.g.*, Bartlett deposition, pp. 105-07 (Ex. P-109).

¹³ Of the nine, all are positive correlations showing some degree of relationship between NTE scores and supervisory ratings with respect to the subject matter knowledge necessary for effective teaching. These are:

County	Subject	Comparison			No. of Cases
		Knowl- edge rating with Common Score	Knowl- edge rating with Area Score	Knowl- edge rating with Composite Score	
Richland	Biology	.520		.536	16
	Elem. Ed.		.274	.293	84/79
Charleston	Media Spec.	.892			5
Colleton	Elem. Ed.		.658		10
	Early Childhood Ed.	.604	.990	.950	13 (Commons) 5 (Area)

The remaining 60 comparisons, although neatly set out in the computer printout, are useless and meaningless. They do not demonstrate that a relationship between the variables exists and they do not demonstrate that it does not exist.

(iii) *The treatment of the data leads to unreliable results.*

Faced with a collection of data on different forms filled out by different raters using different scales, plaintiffs had to summarize the data in a way that, on the whole, makes their statistical comparisons unreliable.

First, plaintiffs apparently selected only six of the 25 forms with which to work.¹⁴ There is no evidence in the record indicating that this selection was other than arbitrary and that use of the data on the remaining forms would not change the conclusions to be drawn.

Second, plaintiffs adjusted all of the ratings by collapsing the broad range of ratings set out on the form into general categories: four categories (personal qualities, professional qualities, teaching techniques and adequate subject matter command) were used for Forms 5, 6; four different categories (instructional skill and classroom management, personal characteristics, professional attitudes and growth and knowledge of subject matter) were used for Forms 1, 3 and 4 (Charleston County); and two different categories (performance statement and teaching area knowledge) were used for Form 4 (Colleton County). Each of these categories contains different numbers and different kinds of items. For example, the item labelled "teaching techniques" applied to Forms 5 and 6 combines eight ratings.¹⁵ Most are different from the seven ratings contained

¹⁴ Form 1 (B-31, Charleston County); Form 3 (B-36, Charleston County); Form 4 (B-39, Charleston County); Form 4 (B-12, Colleton County); Form 5 (B-1, Richland County); Form 6 (B-3, Richland County).

¹⁵ The category labelled "teaching techniques" on Forms 5 and 6 contains ratings as to each of the following:

- Strives for understanding and development of problem solving skills
- Plans and prepares for classroom work
- Uses materials of instruction effectively
- Provides for individual differences
- Communicates well with children
- Requires reasonable standards of neatness, legibility, etc.
- Confers with parents in accordance with needs of children
- Uses disciplinary techniques wisely

B-5

in the item labelled "instructional skills and classroom management" applied to Forms 1 and 3.¹⁶ Form 4 (Charleston County) has a different structure, but it has apparently been made to conform to the general categories used for Forms 1 and 3. Form 4 (Colleton County) has a different structure and only one general category was used to summarize 22 of the 23 ratings on that form.

Third, plaintiffs may have imposed a standard rating scale of 1 through 5 on each of the forms, although most of the forms had not been prepared using that scale. The various scales used on the original forms are as follows:

Form 1	Form 3	Form 4
(no numbers used)	(no word designations used)	(Charleston Co.)
unsatisfactory	1	O—Outstanding
needs improvement	2	P—Professionally Competent
satisfactory	3	I—Improvement Needed
very good	4	U—Unacceptable
superior	5	N—Not applicable

Form 4	Form 5	Form 6
(Colleton Co.)		
1—unobserved	1—unsatisfactory	1—acceptable
2—outstanding	2—below average	2—needs improvement
3—good	3—average	3—unobserved
4—adequate	4—above average	4—not applicable
5—needs development	5—superior	
6—not applicable		

¹⁶ The category labelled "instructional skills and classroom management" on Forms 1 and 3 (B-32, B-37) contains ratings as to each of the following:

- Knowledge of subject matter
- Development and use of effective instructional techniques
- Consistent and careful planning in accordance with authorized courses of study
- Organization and effective direction of pupil activities
- Ability to maintain good class control
- Rapport with pupils
- Classroom environment—ventilation, light, displays, etc.

B-6

The record is unclear as to whether and how plaintiffs changed these ratings to a common (and undefined) scale of 1 through 5 although there is no common denominator by which these scales could be merged, and if they did not change the scale, whether they contend any of the results are comparable.¹⁷

Fourth, plaintiffs calculated composite ratings for each of the composite categories of ratings, thus combining and compounding the factors described above that create unreliability in the comparisons.

After completing all five steps, plaintiffs compared the three NTE scores (Common Examinations, Area Examinations and Composite scores) with each of the composite ratings and calculated correlation coefficients. Few of the correlations are statistically significant. This result is to be expected because most of the comparisons made were between the test scores and qualities or characteristics that the tests make no claim to measure.¹⁸ A similar result would be expected if test scores were compared against the height, weight and hair color of teachers.

(iv) *The comparisons between knowledge attributes and test scores were positive.*

Only one type of comparison made in plaintiffs' study was even arguably relevant. Some rating that had to do with subject matter knowledge was singled out in each of the rating forms and compared with NTE scores. This comparison suffered from all of the infirmities described above. Each of these rating items was different;¹⁹ a

¹⁷ The "record", such as it is, apparently includes only Stipulation No. 11 and Stipulation No. 20 with attachments.

¹⁸ The comparisons for which little or no correlation would be expected are:

Scores on:	against	Ratings for:
Common Examinations		Personal Qualities
Area Examinations		Teaching Performance
Composite of Area and Commons		Instructional Skills and Classroom Management
		Performance Statements
		Professional Competencies and Qualities
		Professional Attitudes and Growth

¹⁹ The formulations varied in specificity and scope:

Forms 1, 3: "Knowledge of subject matter," (B-32, B-37)

[Footnote continued on page B-7]

B-7

different rating scale had been applied to each rating item;²⁰ and most of the sample sizes were very small.²¹ But each of the ratings for which the comparison was statistically significant was *positive*.²² This indicates that, to the extent a criterion-related study produces any relevant evidence, that evidence corroborates the evidence of content validity and suggests that the NTE are a fair measure of the subject matter knowledge required for effective teaching.

¹⁹ [Continued]

Form 4 (Charleston): "Demonstrates wide and growing knowledge in subject area." (B-40).

Form 4 (Colleton): "The teacher is knowledgeable of his teaching area." (B-13).

Forms 5, 6: "Adequate command of subject matter."

²⁰ See p. B-5 *supra*.

²¹ See pp. B-3 to B-4 *supra*.

²² See note 13 *supra*.

APPENDIX C
SUMMARY OF JUDGES' DEPOSITIONS

Name of Judge	Type of Panel	Willingness to Serve	Outside Influence	Timely Selection	Receipt of Instructions	Other Difficulty Identified	Used Best Professional Judgment	Clarity of Instructions CRP	KEP
Adair	CRP	Yes (p.6)	No (p.6)	Yes (p.8)	Yes (p.8)		Yes (p.16)	N/A	
Adams	CRP	No (wasn't particularly happy) (p.9)	No (p.10)	Yes (p.9)	Yes (p.10)		Yes (reasonably sure) (p.27)	Yes (p.20)	
Arnold	KEP	N/A	N/A	N/A	Yes (p.6)		Yes (p.22)		N/A
Barron	KEP	N/A	No (p.12)	N/A	Yes (p.15)		N/A		N/A
Basil	KEP	Yes (p.10)	No (p.17)	Yes (p.10)	Yes (p.16)		N/A		N/A
Bentley	CRP	Yes (p.7)	No (p.9)	Yes (p.7)	Yes (p.9)		Unclear (p.26)	Yes (p.16)	
Berlinghier	KEP	No (annoyance) (p.6)	No (p.7)	Yes (p.6)	Yes (pp.6-7)		Yes (p.22)		Yes (p.16)
Bonnette	CRP	Yes (p.6)	No (p.8)	Yes (p.6)	Yes (p.6)		Unclear (p.26)	Yes (p.15)	
Bopp	KEP	Yes (p.7)	No (p.8)	Yes (p.7)	Yes (p.8)		N/A		N/A

Name of Judge	Type of Panel	Willingness to Serve	Outside Influence	Timely Selection	Receipt of Instructions	Other Difficulty Identified	Used Best Professional Judgment	Clarity of Instructions	CRP	KEP
Boykin	KEP	Yes (p.5)	N/A	N/A	Yes (p.7)		Yes (p.26)	No (p.8)		
Bracey	KEP	Yes (p.9)	No (p.10)	N/A	Yes (p.10)		N/A	No (p.17)		
Brantley	CRP	Yes (p.7)	Unclear (not of a formal nature) (p.9)	N/A	Yes (p.8)		Yes (pp.27-28)	N/A		
Braun	CRP	Yes (p.6)	No (p.8)	Yes (p.6)	Yes (p.6)		Yes (p.26)	Yes (p.14)		
Button	KEP	Yes (p.8)	N/A	Yes (p.8)	No (p.8)	instruct-ions received were incomplete (p.8)	N/A			N/A
Calhoun	CRP	Yes (p.26)	No (p.32)	Yes (p.26)	Yes (p.28)		Yes (relatively) (p.44)	Yes (p.37)		
Cannon	CRP	Yes (p.8)	No (p.13)	Yes (p.8)	Yes (p.12)		Yes (fair) (p.31)	No (pp.30-31)		
Cash	KEP	Yes (p.6)	N/A	N/A	Yes (p.8)		Yes (pp.21-22, 27)			N/A
Caskey	CRP	Yes (p.7)	No (p.6)	N/A	Yes (p.8)		Yes (p.27)	N/A		

Name of Judge	Type of Panel	Willingness to Serve	Outside Influence	Timely Selection	Receipt of Instructions	Other Difficulty Identified	Used Best Professional Judgment	Clarity of Instructions	CRP	KEP
Chewning	CRP	Yes (p.9)	No (p.14)	Yes (p.9)	Unclear (doesn't recall) (pp.10-11)		Yes (higher in some areas than others (p.36)	Yes (p.35)		
Churdar	CRP	Yes (p.8)	No (p.9)	Yes (p.8)	Yes (p.9)		N/A	N/A		
Creel	CRP	N/A	N/A	Yes (p.7)	Yes (p.8)		N/A	Yes (open to interpretation) (p.14)		
Douglass	KEP	Yes (p.7)	N/A	N/A	Yes (p.9)		Yes (p.27)			Yes (but some confusion) (p.28)
Elliott	CRP	Yes (p.12)	N/A	Unclear (doesn't recall) p.13)	Yes (p.12)		N/A	Yes (open to interpretation) (pp.14,20)		
Ellis	KEP	N/A	N/A	N/A	No (p.10)		Yes (p.25)			Yes (pp 10, 11)
Feldman	KEP	Yes (p.8)	N/A	N/A	Yes (p.10)		Yes (p.25)			Yes (p.20)

Name of Judge	Type of Panel	Willingness to Serve	Outside Influence	Timely Selection	Receipt of Instructions	Other Difficulty Identified	Used Best Professional Judgment	Clarity of Instructions CRP	KEP
Finney	CRP	Yes (p.11)	No (p.13)	Yes (p.11)	Yes (p.12)		N/A	No (some) (p.26)	
Foster	KEP	Yes (p.11)	N/A	N/A	Yes (p.13)		N/A		Yes (p.17)
Franklin	CRP	Yes (p.6)	No (p.7)	N/A	Yes (p.7)		N/A	No (p.12)	
Fraser	KEP	Yes (but not enthusias- tic) (p.7)	No (p.9)	Yes (p.7)	Yes (p.8)		N/A		Yes (p.18)
Glover	KEP	N/A	No (p.10)	Yes (p.8)	Yes (p.10)		Yes (p.17)		Yes (open to interpreta- tion) (p.16)
Hall	CRP	Yes (p.74)	No (pp.74- 75)	Yes (p.74)	Yes (p.74)		Yes (p.94)	Yes (subject to interpreta- tion) (pp.82, 84,94)	
Harold	CRP	Yes (p.10)	N/A	N/A	Yes (p.10)		Yes (p.27)	Yes (p.19)	

C-4

Name of Judge	Type of Panel	Willingness to Serve	Outside Influence	Timely Selection	Receipt of Instructions	Other Difficulty Identified	Used Best Professional Judgment	Clarity of Instructions CRP	KEP
Harris	KEP	Yes (p.9)	No (p.10)	Yes (p.9)	Yes (p.9)		N/A		N/A
Hasty	KEP	Yes (pp.9-10)	No (p.12)	Yes (p.10)	Yes (p.11)		Yes (p.21)		Yes (p.17)
Hess	KEP	Yes (p.7)	No (p.10)	Yes (p.7)	Yes (p.8)		N/A		N/A
Hodges	CRP	Yes (p.8)	No (p.10)	Unclear (doesn't recall) (p.8)	Yes (p.9)		N/A	Yes (pp.13-14, 16,17)	
Howe	KEP	Yes (p.8)	No (p.11)	Yes (p.9)	Yes (p.10)	felt rushed (p.25)	Yes (75%-80%) (pp.26,43)		No (vague) (p.13)
Huang	CRP	N/A	N/A	Yes (p.5)	Yes (p.5)		Yes (pp.15,16)	Yes (open to interpretation) (p.12)	
Hudson	KEP	Yes (p.9)	No (p.9)	Yes (p.9)	Yes (p.9)		N/A		N/A
Huger	CRP	Yes (p.5)	No (pp.6-7)	Yes (p.6)	Yes (p.7)		Yes (p.22)	Yes (p.12)	

Name of Judge	Type of Panel	Willingness to Serve	Outside Influence	Timely Selection	Receipt of Instructions	Other Difficulty Identified	Used Best Professional Judgment	Clarity of Instructions	CRP	KEP
James	CRP	Yes (pleased) (pp.8,9)	N/A	No (2 or 3 days) (p.8)	Yes (pp. 9,10, 11)	Not enough time to study; received wrong test. (pp. 9-11,28)	Yes (p. 28)	Yes (pp.18,23)		
Jeremias	KEP	Yes (p. 4)	N/A	Yes (p.5)	Yes (pp.4,5)		Yes (p. 32)		Yes (p. 24)	
Kimball	CRP	Yes (p.9)	No (p.13)	Yes (p.10)	Yes (p.11)		Yes (p.41)	Yes (p.33)		
King	CRP	Yes (p.11)	No (pp.12, 13)	Yes (p.11)	Yes (p.11)	Was given wrong area to review. Was not told he would be doing chemistry/physics (p.30)	No (p.17)	No (pp.18, 30)		
Luckawecki	CRP	Yes (pp.7-8, 29)	N/A	N/A	N/A		Yes (pp.25,26, 27,34)	Yes (p.29)		

C-6

Name of Judge	Type of Panel	Willingness to Serve	Outside Influence	Timely Selection	Receipt of Instructions	Other Difficulty Identified	Used Best Professional Judgment	Clarity of Instructions	CRP	KEP
McFadden	KEP	Yes (p.9)	N/A	Unclear (did not know when) (p.9)	No (p.10)	Questions were not answered, didn't get pertinent responses, ignored. (p.11)	N/A	No (pp.11, 12, 13, 18, 25, 26)		
D.Mercer	KEP	Yes (p.8)	No (p.10)	Yes (p.8)	No (p.9-10)		Yes (p.22)	Yes (pp.22, 33)		
E.E.Mercer	CRP	No (it was a chore) (pp.8,9)	No (pp.12, 13)	Yes (p.9)	Yes (p.11)		Yes (p.9)	Yes (p.19)		
Mitchell	CRP	Yes (p.16)	N/A	Yes (p.17)	No (but not certain) (pp.17,18)		Yes (pp.40-41)	Yes (pp.23, 26, 35, 36)		
Morgan	CRP	Yes (but mixed reaction) (p.10)	No (p.12)	Yes (p.10)	Yes (p.11)		Yes (pp.35, 46)	No (pp.21, 22, 26, 30, 31, 44)		
Nicholson	KEP	Yes (p.9)	No (p.13)	Yes (p.9)	Yes (p.11)		Yes (p.36)	Yes (but open to interpretation) (pp.18, 34-35)		

C-7

Name of Judge	Type of Panel	Willingness to Serve	Outside Influence	Timely Selection	Receipt of Instructions	Other Difficulty Identified	Used Best Professional Judgment	Clarity of Instructions CRP	KEP
Park	KEP	Yes (pp.8-9, 10)	No (p.10)	Yes (p.9)	Yes (p.10)		Yes (p.28)		No (open to interpretation) (p.27)
Paysinger	KEP	Yes (p.7)	No (p.8)	Unclear (does not know) (p.7)	No (can't remember getting anything) (pp.7,8)		Yes (pp.10,15)		Yes (pp.9, 15)
Platt	CRP	Yes (p.5)	No (pp.7,8)	N/A	Yes (p.6)		Yes (pp.32-33)	Yes (p.16)	
Pollard	CRP	Yes (p.102)	N/A	Unclear (does not remember) (p.102)	Yes (not sure it was instructions) (p.102)		Yes (pp.118, 127)	Yes (pp.109, 124)	
Powell	KEP	Yes (p.6)	N/A	N/A	Unclear (not sure) (p.9)	projector problems; seating arrangements. (p.13)	Yes (pp.16-17)		Yes (p.40)
Price	CRP	Yes (p.10)	No (p.13)	Unclear (does not remember) (p.11)	Yes (pp.11,13, 20)		Yes (p.42)	N/A	

Name of Judge	Type of Panel	Willingness to Serve	Outside Influence	Timely Selection	Receipt of Instructions	Other Difficulty Identified	Used Best Professional Judgment	Clarity of Instructions CRP	KEP
Redd	KEP	Yes (pp.5,6)	No (p.7)	Yes (p.5)	Yes (p.6)		Yes (p.19)		N/A
Redderson	KEP	Yes (pp.6,24-25)	No (p.8)	Yes (p.6)	Yes (not sure what contents were) (pp. 7,25)		Yes (p.27)		Yes (pp.11-12,26)
Reitzer	CRP	Yes (although not entirely clear) (p.8)	No (pp.9,10)	Yes (p.8)	Yes (p.9)		Yes (p.37)	Yes (pp.17,20,35)	
Robinson	KEP	Yes (p.7)	N/A	N/A	Yes (p.13)		N/A		Yes (p.19)
Ropp	CRP	Yes (p.7,43)	No (pp.11-12)	Yes (p.8)	Yes (pp.8,43)		Unclear felt qualified (p.43)	Yes (pp.17,18,44,45)	
Rouse	KEP	Yes (p.8)	No (p.10)	Yes (p.9)	Yes (p.8,10)	Should have stayed home (p.9)	Yes (p.21)		Yes (p.20)
Sargent	CRP	Yes (p.8)	N/A	N/A	Yes (p.8)		N/A	N/A	
Sharpe	CRP	Yes (p.7)	No (p.9)	Yes (p.8)	Yes (p.3)		Yes (p.32)	Yes (pp.13-14,22,31)	

C-9

Name of Judge	Type of Panel	Willingness to Serve	Outside Influence	Timely Selection	Receipt of Instructions	Other Difficulty Identified	Used Best Professional Judgment	Clarity of Instructions CRP KEP
Sherard	KEP	Yes (p.10-11)	No (p.13)	Unclear (sometime in October) (p.12)	Yes (p.11)		Yes (p.29)	N/A
Silvernail	KEP	Yes (p.9)	No (p.11)	Yes (p.9)	Yes (p.9)		Yes (pp.29-30)	Yes (p.29)
Sins	KEP	Yes (p.12)	N/A	Unclear (does not remember) (p.12)	Yes (p.17)		No (p.30)	Yes (p.31)
Solomon	CRP	Yes (p.10)	No (p.14)	Yes (p.10)	Yes (p.13)		N/A	No (had some problems) (pp.17,19,24,29)
Spove	CRP	Yes (p.5)	N/A	Yes (p.6)	Yes (p.6)		Yes (pp.23,30)	Yes (pp.25,30)
Staley	KEP	Yes (p.10)	No (p.11)	Yes (p.10)	Yes (p.11)		N/A	No (p.21)
Stauss	KEP	Yes (p.6)	No (p.8)	Yes (p.6)	Yes (p.7)		Unclear (yes & no) (pp.14-16)	Yes (p.13)

Name of Judge	Type of Panel	Willingness to Serve	Outside Influence	Timely Selection	Receipt of Instructions	Other Difficulty Identified	Used Best Professional Judgment	Clarity of Instructions	KEP
Swailes	CRP	Yes (p.134)	No (p.136)	Yes (p.134)	Yes (p.134)	Wrong tests given to some individuals. (p.147)	No (on some) (pp.141, 147, 148, 149)	No (p.147)	
Tribble	KEP	Yes (p.7)	No (p.8)	Yes (p.7)	Yes (p.8)	Task to perform was difficult pp.15-16)	Yes (pp.17, 23)		N/A
Van Doren	CRP	Yes (p.10)	No (p.14)	Unclear (cannot remember) (p.11)	Yes (p.10, 13)		No (not too confident) (p.27)	Yes (p.25-26, 27)	
Vinson	CRP	Yes (could have re-fused) (p.5)	No (p.7)	Yes (p.6)	Yes (p.6)		Yes (pp.26-27)	Yes (p.12)	
Von Kleinst	CRP	Yes (pp.5, 35)	N/A	Yes (p.5)	Yes (p.6)		Yes (p.42)	Yes (pp.16, 26, 37)	
Wagner	KEP	Yes (p.7)	No (p.10)	Yes (p.8)	Yes (p.9)		Yes (p.31)	Yes (p.30)	
Waterfall	KEP	Yes (pp.7, 25)	No (p.8)	Unclear (does not remember) (p.7)	Yes (p.8)		Yes (did it fairly) (pp.15, 27)		Yes (pp.25-26, 28)

Name of Judge	Type of Panel	Willingness to Serve	Outside Influence	Timely Selection	Receipt of Instructions	Other Difficulty Identified	Used Best Professional Judgment	Clarity of Instructions	CRP	KEP
Whitaker	KEP	Yes (p.6)	N/A	Yes (p.6)	Yes (p.7)		Yes (p.13)			Yes (p.14-15)
Wiggins	CRP	Yes (p.14)	Unclear	Yes (does not recall exactly; sizeable period) (pp.15, 17)	Yes (p.15)		Unclear	Yes (p.38)		

TOTAL

Asked	75	Asked	56	Asked	63	Asked	80	Participating	81	Asked	60	CRP mem- bers	41	KEP mem- bers	40
Not ask.	6	Not ask.	25	Not ask.	1	Not ask.	1			Not ask.	21				
	81		81	ask.	18		81				81				
Of those asked:		Of those asked:		Of those asked:		Of those asked:		Found difficulty	9	Of those asked:		Asked	35	Asked	28
												Not ask.	6	Not ask.	12
													41		40
Yes	72	Yes	0	Yes	52	Yes	72	No comment	2	Yes	50	Of those asked:		Of those asked:	
No	3	No	54	No	1	No	6		70	No	4				
	75	Unclear	2	Unclear	56	Unclear	2		81	Unclear	5				
				clear	10		80				59	Yes*	27	Yes*	22
					63							No	7	No	6
												Unclear	1		28
													35		

195

* Some felt the questions were open to interpretation.